UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
Facoltà di Scienze Matematiche Fisiche e Naturali
Corso di Laurea Magistrale in Informatica

Virtual social interactions in an affective driven environment

Supervisor: Prof. Domenico Giorgio SORRENTI

First co-supervisor: Prof. Giuseppe BOCCIGNONE

Second co-supervisor: Dott. Simone BASSIS

Master Thesis of:
Jonathan VITALE
Matr. 744539

Accademic year 2011-2012

Table of Contents

Introduction

Aim of this work is to investigate a system able to detect facial expressions and to use them in a model for automatic affect recognition, in order to further investigate models for social interactions mediated by social signals.

Human computer interaction has undergone a great change during the last decades. Currently, thanks to new methods and technologies, we are able to give to the user the possibility to interact with systems simply using gestures and motion.

A lot of applications in disparate fields are going into this direction, in particular videogames field. Several videogames propose an “intelligent” interaction with a virtual avatar using gestures or particular joysticks, however these interactions are not really sociable, as the avatar cannot understand the emotional state of the user and it is not able to establish believable social interactions.

Humans are experts in social interactions. Therefore technologies able to adhere to the social expectations of the people, for example making use of the affective state of the user, are crucial in order to further improve the user experience in our everyday appliances [4]. To achieve this purpose, Affective Computing offers several techniques to extract the emotional state of the user from visual, audio and bio signals, with more or less precision and fortune [5, 10]; nevertheless in recent years several *emotional applications* born from disparate laboratories around the world.

Thanks to the several theories of emotions enounced in current and past years, it is possible to have a useful guideline for the development of an affect recognition model. In this work, the theory of the core affect of Russell [41] plays a crucial role, since it allows to describe emotions as a set of latent variables, which defines a core affect space where the points describe all the possible emotional states of the subject. Conversely, other theories describe the emotions using a limited number of classes or labels. This labels own a too wide meaning, since they are created in a further step by human cognition in order to only help us to classify things. As with the colours, which are classified into classes by humans (red, blue, yellow, green, ...) but indeed their visible hue depends on the wavelength of the light source, emotions can be classified into classes (sadness, happiness, anger, ...) but their exact affective value depends on a small set of latent affective variables.

Current researches on affect recognition [29] state that the observation owns an affective power describable in terms of affective variables. This in turn allows to regress a space on the basis of these specific set of chosen affective variables. In our vision the behaviour is exactly the opposite: it is the latent space to possess an affective power and to generate an observation starting from a specific point of the latent space, which maximizes the likelihood w.r.t. the observation under exam. Consequently this generated observation allows the subject to understand

the affective state of the partner making comparison with the affective character suggested by the point of the latent space, from which this observation is generated. This is for example one of the theories endorsed by mirror neurons theorists.

In this way it is possible to investigate the topology of the latent space, without considering any kind of label on the elements of the database. This in turn implies that the topology of the latent space can be generated considering only the raw observations through unsupervised techniques of machine learning, for instance techniques of dimensionality reduction.

For this reason in this work is proposed a non-linear dimensionality reduction model based on Gaussian Processes, namely the Gaussian Process Latent Variable Model [57].

In order to capture facial expressions from video streams necessary to train the model, an architecture of face detection and normalization is proposed. This architecture involves a face detector [66] and a tracker based on Kalman filter [69]. Furthermore, a facial landmark detector [76] is used to discover angles of rotation of the face and consequently its normalization. Finally, a procedure for light conditions normalization is used in order to remove noise due to ununiform light conditions [79].

The thesis includes some preliminary experiments to test the classification accuracy of the model proposed. The tests are made using two kinds of facial expressions datasets: a dataset based on videos collected from the Web with no labels available, and a dataset created in a laboratory asset presenting several combinations of Action Units used to objectively describe the current facial configuration [81].

For the first dataset only a qualitative evaluation is proposed, for the reason that without labels it is not easy to produce objective and numerical results. However this first test allows to understand the behaviour of the model proposed and its first advantages and defects.

For the second dataset an objective numerical evaluation is used to better interpret the results. Also in this case it is difficult to produce sound results. This is due to the fact that several issues do not allow an accurate temporal alignment of data used for the final comparison. Nevertheless, these results permit a first guideline for future improvements.

The thesis is organized as follows:

Chapter 1 describes the general domain in which our work is placed and develops the general idea that leads us to investigate the issues presented above;

Chapter 2 provides a detailed review of emotions and theories of emotions developed during the past years in order to have an accurate picture of this domain;

Chapter 3 discusses the model chosen for our purpose of emotions recognition and introduces its theoretical foundations;

Chapter 4 explains how it is possible to extract faces from videos and describes our architecture built to fulfil this task;

Chapter 5 illustrates the tests made with our model and the relative results;

Chapter 6 presents our conclusions and the future possible improvements.

Chapter 1

General overview

> A new idea is delicate. It can be killed by a sneer or a yawn; it can be stabbed to death by a quip and worried to death by a frown on the right man's brow.
>
> Ovid

1.1 A social way of interaction

Each instant of our life is a constant interaction with what we call *reality*. We can see because the cells inside our eyes are stimulated by light stimuli; we can touch thanks to the several nerve endings under our skin; we can hear, taste, smell, feel pain ... all using our sense organs. A combination of these sensorial experiences produces the reality in which we are surrounded.

For this reason we are able to interact with our appliances and use these for the everyday life. However, if we look at the world of computers and digital environments there is the necessity to create interfaces that allow communication and interaction between the real world and the digital one. There are available several kinds of interfaces, and each one of these produces a model of interaction. Then, what kind of interface and consequently model of interaction is the best that fits user's needs?

Answering this question is not a so easy task: to prove it there exist several disciplines such as Human-Computer Interaction, Psychology, Design, etc... that try together to respond this question since several decades without giving a unique answer. However we can consider a fact. Humans are social animals and for this reason they spend most of their time interacting with one another. For us it is easy to communicate and interact with someone using natural language, prosody, facial expressions, gesture and so on, for the reason that we grew up through these forms of *natural interactions* and we exploited them along learning.

Accordingly, these modes of interaction have the advantage of being usual, comfortable and to enhance affordance, so they are preferable in most situations. Moreover, unlike non-natural user interfaces, *natural user interfaces* allow to extract some important cues to use into applications for increasing their value to the user and meet more consumers' needs.

At this point it is necessary to introduce and define the term *anthropomorphism*. In general this word refers to the tendency to attribute humanlike characteristics, intentions, and behaviour to nonhuman objects [1]. Psychological studies suggest that increasing accessibility to the human schema[1] results in anthropomorphism [2], that means for example that an object (real or digital) imitating human behaviours or appearance increases accessibility of human schema, and consequently increase anthropomorphism. Obviously is not necessary to anthropomorphize an object in order to use natural interactions on it, in fact to fulfil this purpose we need only a set of natural user interfaces, however if we want to create a greater illusion of a *social interaction* between the user and the real or digital thing, it is crucial to give an anthropomorphic vision of the artefact.

This practice is necessary in order to create a *metaphor* that allows people to use interfaces taking advantage of previous cognitive model learned with the aim to fulfil tasks related to metaphor itself. If an object is endowed with anthropomorphic characteristics it is more likely that people interact with it by using modes of interaction analogous to those used for human-human interaction in a social domain (Fig. 1.1).

The use of metaphors is a fundamental part of our reasoning. A metaphor is defined as a mapping from a set of correspondences between a source domain to a target domain. These correspondences allow us to reason in the target domain using the knowledge we have on the source domain. For example we use the knowledge about the classical mail to create an useful mapping to electronic mail domain [3].

Fig. 1.1: Asimo, an example of anthropomorphic robot interacting with people

Indeed, humans are experts in social interactions. Therefore, if technology adheres to the social expectations of the users, users will find the interaction enjoyable and they will experience stronger feelings more congruent with their expectations. It has been shown [4] that people prefer to interact with machines in the same manner in which they interact with other people. Thus, it is useful to study the implementation of a metaphor that maps the correspondences from the domain of the machine to that of a common social interaction. This is the purpose of this work, as it will be explained in later sections.

[1]Set of characteristics, models and behaviours strictly related to humans

1.2 What is Affective Computing?

Affective Computing is a quite new research area at the intersection between Psychology and Computer Science, originated with Rosalind W. Picard [5] at MIT, who framed it as follows:

> *Computing that relates to, arises from or deliberately influences emotions.*

As it can be seen by the definition, the role of emotions in this field is crucial. However, why a research theme typical of Psychology could be important for a research area of Computer Science?

Actually there are several cases of psychological themes that were investigated by computer scientists, but it is interesting how emotions, that in the mainstream of Western culture is deemed to be ruled by irrational processes, could be of interest in a field traditionally governed by logic, determinism and rationality. To better understand this strange combination, consider the following investigations about emotions and their relation with perception and cognition.

A study by Cytowic [6] has investigated synesthetic experience of individuals. A synesthetic experience consists in associating an experience felt with a sense organ, with another feeling typical of a different sense organ, for instance "seeing" colours while hearing music. Cytowic investigated the behaviour of the cortex during a synesthetic episode. As result, an overall increase of brain metabolism occurred in the limbic system, and not in the higher cortex, where it was expected. The limbic system has traditionally been assumed as the set of brain regions supporting emotion, memory and attention (although the very concept of limbic system has been recently questioned, cfr. LeDoux [7] for a deeper discussion). Its activity during synesthesia shows that the limbic system has a crucial role in perception. Indeed, it is common to perceive the world around us on the basis of our emotional state, for instance the reality it could be see *through rose-coloured glasses* during a joyful state.

Findings from recent studies have provided evidence that no clear cut can be drawn between emotion and perception (and more generally cognition, see Pessoa [8] for an in-depth discussion). It is worth recalling the important study by Damasio [9]. Damasio's patients have injuries in the part of the cortex (orbitofrontal cortex, OFC) that communicates with the amygdala (a region of the limbic system). These lesions involve the inability to regulate interactions between the emotional responses and cortical decision-making structures. Thus, Damasio's patients *appear* to be intelligent and very rational, but they are actually unable to make decisions or they spend too much time compared to healthy subjects, also if the decision to take is simple. Damasio supports the hypothesis that emotions regulate decision-making as a necessary bias (the Somatic Marker Hypothesis, SMH) to evaluate potential outcomes, and prevent an infinite logical search.

The implications of such findings are significant also for computer science and industry: computers, if they are to be truly effective at decision-making, will have to be endowed with emotion-like mechanisms working in concert with their rule-based systems [5].

1.2.1 Domains and applications

Like most areas of computer science, Affective Computing can be used in a wide range of domains and applications, relying on sensing and recognizing user emotion [10], or on generating expressive affective behaviours in synthetic agents and robots [11].

One remarkable example is the *entertainment* domain, and a possible application concerns videogames (*Affective Gaming*). This new form of videogames exploits the affective state of the user to calibrate and to manage the game difficulty and/or the gameplay. Current focus in Affective Gaming is primarily on the sensing and recognition of the players' emotions, and on tailoring the game responses to these emotions; e.g., minimizing frustration, ensuring appropriate challenge [12].

There are several reasons to use emotions as a way of input in videogames. For instance the game could adapt history and events on the basis of the affective state of the player. An example of this feature could be a loud noise in an horror based game when the player is incredibly tense to augment the fear of the user and give to it a best game experience [13].

More generally affective computing could be crucial in the future of human-computer interaction. As discussed in Section 1.1, humans have a bias to threat things as people. By creating new forms of interaction based on emotions it is possible to enhance the overall quality of user experience. An example is provided by *embodied conversational agents* (ECA), namely interfaces governed by virtual or robotic agents that express and recognize the affective state of the user and use this cue to help the user during the interaction with the application (see Isbister [14] for more information). A primordial example of ECA is the Windows Office assistant Clippy, although it was not able to use or express emotions.

Another domain of interest is represented by "technologies as means of persuasion" that aim at changing the behaviour, feelings and attitudes of users [15]. The *health domain* is one among the most interesting fields where Affective Computing can be effectively applied to improve user's security and health. Robots or virtual avatars can be used in therapies, such as the treatment of autism [16]. Other applications can be those concerning elderly people, especially those who live alone [17].

No less important are applications concerning *decision-making* and *security*. As previously mentioned, emotions could be crucial to improve the quality of decisions in an intelligent artificial system. Further, the use of emotions recognition could be central to those applications involved in security, for example to detect acts of violence or as tutoring systems to prevent *loss of situation awareness* by the user along critical episodes [18].

Many other possible domains and applications can be devised, where Affective Computing might play a role: the limit only is in the creativity and fantasy of researchers.

1.3 The concept

1.3.1 A natural learning process

A disapproving glance can turn us in a bad mood, while when we are praised, we feel positive. Following Damasio [9], negative feelings will prevent the individual from falling back in disagreeable (mental and physical) situations, as opposed to positive feelings that associate a given event with a profitable outcome [19]. On such basis, emotions are able to regulate the learning process and decision-making of individuals.

There has been a time when we were babies and we did not know exactly the meaning of words. Our parents helped us to discover the world around us by indicating things and suggesting the appropriate names. They spoke with a particular intonation and acted facial expressions so to suggest the emotional valence of the scrutinized objects. This kind of progressive learning was not limited to things, but also extended to events and behaviours, so that we could learn the right way to act [20].

The same process of learning happens to pets too. When a new puppy comes in our house and for example destroys our shoes, we severely rebuke it to avoid the same behaviour in the future. Infants and puppies are not the only cases of this form of learning: for instance, when adults fail at work, their boss warning might trigger embarrassment useful to prevent from the same error in the future.

This form of learning requires a kind of social interaction based on emotional signals and it is crucial investigating a model for emotions representation and relative dynamics during several social interactions. Next sections will try to better illustrate the overall idea.

1.3.2 An intelligent emotional avatar

Drawing inspiration from this common process of learning, we propose to create a virtual avatar able to recognize the affective states of users in order to learn new actions and behaviours. To fulfil this purpose, a camera (eventually, a Kinect[2] camera) and a microphone will be available in order to capture frames of the user's face, and cues of the user's speech.

The avatar should be able to express emotions through facial expressions that derive from avatar's internal affective state, which is in turn regulated by the social interactions with the users. This behaviour is necessary as a feedback for the user.

At the beginning the avatar will be endowed with a small subset of basic behaviours. One of these will be the ability to mimic and eventually learn the actions of the user that, in this case, will be a new available action that the avatar could select in the future. On the basis of the affective state of the avatar a specific behaviour will be taken and obviously the user could punish or reward the avatar, enabling him to recognize what kind of behaviour is better to take under specific circumstances (Fig. 1.2).

[2] www.kinectforwindows.org/

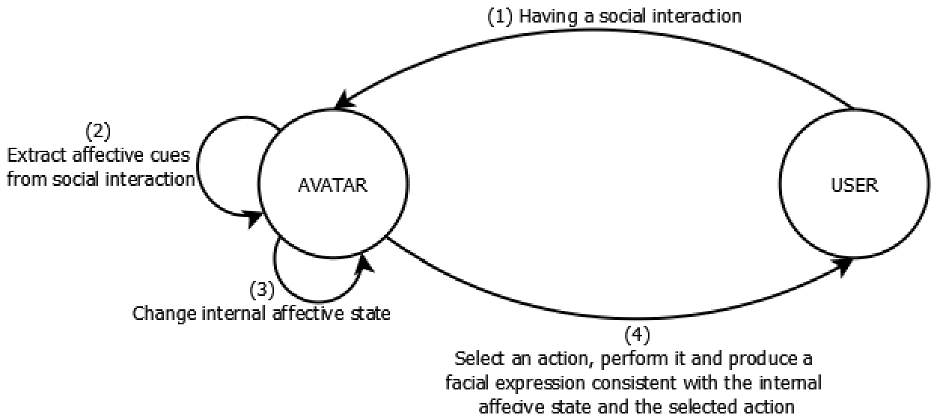


Fig. 1.2: Model of interaction between the avatar and the user

1.3.3 Previous works

At a first sight this project could appear like science fiction, however some important steps have already been taken in this direction. Important examples of research projects are *Kismet* and a more sophisticated version, *Leonardo*, from the MIT Media Lab [21, 22].

Kismet is a social robot able to communicate its internal emotive state by emulating the basic emotions of humans. It can communicate its emotive state through facial expressions, body posture, gaze direction and quality of voice [23]. Kismet's facial expressions are generated over a three-dimensional affect space, where each point (or cluster of points) of this space governs a specific facial expression, interpolating it along neighbouring points.

The three dimensions correspond to arousal (high/low), valence (good/bad), and stance (advance/withdraw) [23] and are inspired to the *theory of core affect* of Russell (see Section 2.2.6). As core affect theory claims, the current affective state of the robot is represented by a single point of this space. The dynamic of this point along the three axes of the space, involves the change of its facial expression on the basis of the trajectory of the point representing its internal affective state.

The affective state of Kismet can change with an interaction with it. It could for example get bored if the interaction is not so exciting, or get surprised if a particular interesting object is shown to it. However Kismet has no real cognition of the surrounding world and most of its interactions with the world are low-level processes based on more or less sophisticated saliency map representations. Unfortunately, the cognition of things and persons, the theory of mind, the recognition of self and other broader questions are issues very complicated to cope with, and for this reason the major efforts of Kismet's creators are on human-robot interaction leaving open, at least for now, issues most related to artificial intelligence.

Leonardo is the evolution of Kismet, and for this reason, it has more powerful capacities of expression and it is able to learn through imitation and spatial scaffolding. It has 64 degree of freedom and, unlike Kismet, it has a complete humanoid body. The design is targeted for rich social exchanges with humans as well as physical interactions with the environment [24]. It is able to communicate with gestures and facial expressions to people and it has the ability to manipulate objects.

The robot can locate and identify the facial features of a human partner by using a camera and a software of facial features tracking. Thus, Leonardo is able to perceive a scene and understand what it is currently happening using a tree structure where each leaf of the tree is specialised to extract a specific feature, like faces.

An action system is responsible for behaviour arbitration of the robot, drives it through a decision-making process and instructs the motor system on how physically implement the action selected.

The learning process occurs through an imitative interaction inspired by developmental psychology theories. There are two phases: the first one consists of the imitation of Leonardo's facial expression by humans, and the second one where Leonardo imitates human facial expression. This continuous cyclic process leads to an imitative learning of facial expressions through a form of emphatic consciousness.

There are many similarities between the works of Breazeal and the work described here. Clearly, it is impossible to consider all the aspects and features quoted above in a single thesis work. For this reason, we will begin to concentrate on the first step of the project that will be the basis for future work.

Here, we will mainly consider aspects concerning the detection and modelling of facial expressions, as a first step for investigating and design models of affective social interaction.

As usual in most of computer science's studies, the main task of a computer scientist is to model a complex and real problem in a simple and tractable one, eventually with some simplifications. This task is often accomplished using mathematical models.

The first part of the work investigates machine learning algorithms able to generate a model enough informative for our purpose. Methods of dimensionality reduction will be essential to make the problem tractable in term of time and space complexity.

The second part of the project detection of faces and the consequently extraction of important cues about facial expressions will be crucial. In this effort, computer vision techniques and image processing algorithms play a central role.

Chapter 2

A world of emotions

> Human behaviour flows from three main sources: desire, emotion, and knowledge.
>
> Plato

2.1 What is an emotion?

It is not so simple to answer the original question posed by James [25]. If someone asks to us to define what is an emotion, probably we will be in serious difficulties and we will describe the concept by examples. Psychology researchers, during these past decades, tried to understand what an emotion is, but at present we have only several theories that often are in conflict with each other.

Our brain is very complex and, though neurosciences have made great stride in recent years, we do not have sufficient knowledge about its behaviour yet. Nevertheless we have some important cues that allow us to exclude some theories and to make plausible others. Anyway, what is an emotion, and why it is so difficult to give a clear definition?

Emotion is a complex set of interactions among subjective and objective factors, mediated by neural hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labelling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behaviour that is often, but not always, expressive, goal-directed, and adaptive [26].

Anyway, providing a single, simple and clear definition is not an easy task, and probably it is not even feasible or correct.

In the history of Western thought and more specifically in the most recent science of affects, different approaches have addressed different aspects of such a complex phenomenon, hence giving rise to different definitions.

These main theories of emotions will be described in the next section.

2.2 Main theories of emotions

Several theories were born during last decades to try to define what is an emotion and to understand the mental processes that take place during an affective process.

In this way, the concept of emotion is expanded to the concept of "emotive episode", from which can be extracted several components. Among these there are for instance a cognitive component, a sensational component (emotive experience), a motivational component (actions tendency), a somatic component (physiological responses) and a motor component (expressive behaviours of the emotion) [27].

Each theory has a different interpretation of the several single components during an emotive process. These differences can be either the number of components identified in an emotive episode or the definition of emotion itself defined as one or more components of the emotive episode.

Other differences among theories concern the representation of emotions, which can be classified into a discrete number of classes or represented by point in a multidimensional space, where discrete classes are nothing more than a cluster of points. Inside each of these two strands we have debates concerning, for example, the number of classes and their labels or the number of axes and the relative labels of the space's variables.

To better understand and analyse a theory of an emotional process, Houwer suggests a series of questions that should be addressed by theories [27]. The first question (Q1), named "*elicitation problem*", aims at understanding which stimuli of the environment causes an emotion and which does not. This problem includes two subproblems: the first (Q1A) asks which stimuli produce an emotion and which not, the second (Q1B) asks how the organism performs this task.

The other two questions concern the quantitative and qualitative aspects of emotions. The quantitative aspect is known as "*intensity problem*" (Q2) and include two subproblems: the first (Q2A) asks which stimuli cause strong emotions and which cause weak ones; the second (Q2B) investigates the mechanisms that determine the intensity of emotions. The qualitative aspect is known as "*differentiation problem*" (Q3) and include two subproblems: the first (Q3A) asks which stimuli cause positive emotions and which cause negative ones; the second (Q3B) investigates the mechanism that determine the qualitative aspect of an emotion.

2.2.1 James' theory

James is known to have changed, with his theory [25], the order of the events in an emotive episode. For James it is not the emotional experience that activates facial muscles and other physiological responses, but are the physiological responses that make the subject feel an emotive experience ("*perception of bodily changes is the emotion*", [25]). That means, for instance, that we do not tremble because we have fear, but we have fear because we tremble and with this physiological behaviour is associated the emotional experience of fear.

Both intensity and quality of emotions are determined by the intensity and quality of the physiological responses produced after the stimulus. For this reason James' theory give an answer to the intensity and differentiation problem, but not to the

elicitation problem, namely it does not specify how the physiological responses are produced after a stimulus.

This theory was highly criticized at the time and more recent investigations and experiments have provided controversial evidence.

However, the idea of emotion as embodiment is a long lasting trend and a more sophisticated modern and revised version is that proposed, based on recent neurobiology findings, by Damasio [28].

Most important from the Affective Computing perspective, is that under the rationale that emotional experience is embodied in peripheral physiology, systems can detect emotions by analysing the pattern of physiological changes associated with each emotion (assuming a prototypical physiological response for each emotion exists). The amount of information that the physiological signals can provide is every day increasing, mainly due to major improvements in the accuracy of psychophysiology equipment and associated data analysis techniques [29].

2.2.2 Affect program theory

Close to the emotion as embodiment approach, one can find the evolutionary or affect program approach [30].

The aim of affect program theories is not to explain the process of an emotive episode, but to explain how a stimulus is able to cause the effects of a particular emotion selected during an emotive episode.

This theory supports the idea that, during evolution, were created dedicated neural circuits for each of the six basic emotions. If the activation of the specific neural circuit pass the threshold, a *program* of the selected circuit will be activated, that is physiological signs, tendency of actions and emotional feelings are manifested in the subject.

Basic emotion theories, inspired by Tomkins [31] rediscovery of Darwin's [32] work on the expression of emotion, were developed by Ekman [33] and Izard [34] (cfr., Section 2.4 below).

2.2.3 Schachter's theory and the shift towards cognition

Schachter [35] tries to solve the critics and issues of the original James' theory. He states that a stimulus has the ability to produce physiological responses in a subject and that responses are then interpreted in a next step by a cognitive process. This in turn identifies the particular emotion and consequently displays it in the subject ("perceived arousal leads to labelling feelings as an emotion based on situational cue").

In this way, the physiological responses are not the only causes of an emotive state, but it is involved a cognitive process too. For instance, a dog that dazzle to us and meeting our boyfriend can cause the same level of arousal, however it is only at the level of cognitive process that we attribute in the first case a danger rather than joy in the second case.

Nevertheless, this theory is not able, as well as James' theory, to answer the elicitation problem; in fact, after the physiological stimulation phase, it considers as

main component a not well defined cognitive process.

2.2.4 The cognitive approach: Appraisal theory

Appraisal theories [36, 37, 38] are at the core of a cognitive approach to emotions and are usually opposed to the emotion as embodiment approaches.

This theory supported by several psychologists in the course of several years, argues that cognition is antecedent of emotion, as well as Schachter did, however it does not give to this cognitive process a conscious factor, but an automatic or unconscious one.

This idea born after critics moved by Zajonc [39] to Schachter's idea, who shows how it is not necessary a conscious cognitive process to display an emotion; nevertheless the data presented by Zajonc do not demonstrate the inexistence of an unconscious cognitive process.

Another difference with Schachter's theory is that appraisal theory inserts the cognitive components immediately after the stimulus and before the physiological response of the subject. In this way it is the cognitive components to have the task of establishing what kinds of physiological responses cause to the subject, on the basis of the stimulus characteristics and so answering to the elicitation problem (Q1). Furthermore, is the cognitive component again to determine quality (Q3) and intensity (Q2) of the emotion.

After the emotional experience of the subject, it is put another cognitive process, however this one is conscious and it serves to attribute the correct emotional label. This means that it is the subject itself to determine an emotional label after feeling an emotive episode; however it is the unconscious and automatic cognitive process that determines the physiological behaviour of the subject after a stimulation.

Emotion researchers supporting this theory tried to understand which kind of stimuli produce emotions and which do not. Nevertheless it is difficult –if not impossible– to determine one-to-one relationships between stimuli and emotions, because they depend on the subject's beliefs and on the context.

Anyway, it emerged that these stimuli could be identified by variables and these variables could be easily classified in such a way that they help to determine the intensity and quality of an emotion. Each of these variables considers a particular aspect of the stimulus. A set of these values creates an appraisal pattern that, by assumption, is in relationship with one particular emotion. An example of appraisal variable is the *relevance to the goal*.

Summing up, these theories assume an emotion architecture that is based on an individual subjective evaluation or appraisal of the significance of events for their wellbeing and goal achievement, postulating a specific set of appraisal criteria (e.g. novelty, intrinsic pleasantness, goal conduciveness or motive consistency, agency, responsibility, coping, legitimacy and compatibility with self and societal standards).

2.2.5 The cognitive approach: Network theory

Bower's NetworkTheory of Affect [40] is another variant of a cognitive approach to emotion. It supports the idea that emotions are coded into memory and the acti-

vation of these memories is the main cause of emotion (Q1). At the beginning only few relevant stimuli cause emotions, then these stimuli are progressively processed through conditioning procedure.

When a new emotive episode occurs, the information about the stimulus and the related physiological responses are coded into memory and, through a continuous matching of the stimulus with another one previously coded into memory, it takes the same emotional valence of the previously learned stimulus. If the new stimulus does not match with others stimuli of different schemas, a new schema will be matched with a generalization process.

This theory answers the elicitation (Q1) and differentiation (Q3) problem by adding a cognitive component, which uses the schemas coded into memory to match a new stimulus into the most suitable of these. The intensity of the emotion (Q2) depends on the intensity of the activation of the schema selected by the cognitive component.

2.2.6 Theory of the core affect of Russell and of conceptual act of Barrett

With respect to previous approaches, Russell's aim [41] is to provide a synthesis under the rationale that previous and competing approaches basically addressed different.

In this perspective, the idea that emotions can be reduced into a limited number of classes, it is not realistic. Russell supports the idea that these categories are not given by nature, but are artificial constructions made by society and culture. On the contrary Russell claims the existence of emotive variables of valence and arousal, which are the real basic constituents of emotional life: *"continuous core affect constituted by valence and arousal is interpreted and categorized in the light of situational cues"* [41].

It is possible to do a comparison between emotions and colours. Society and culture have categorised colours into classes like red, blue, yellow... however we know that there are a lot of shading inside a single category and all depends on the wave length of light stimulus, which is a continuous variable. At the same way, emotions are categorised into classes like anger, happiness, joy... but the real constituents of these are two continuous variables of valence and arousal.

These variables are defined as property of stimuli, of the neurophysiological states and of the conscious experiences. The combination of both valence and arousal values it is called "*affective quality*".

The affective quality of a stimulus cause in the subject a state called "*core affect*" with consequences both neurophysiological and mental.

Barrett [42] agrees with Russell's vision, however she tried to understand how these points of affective space can be categorised into classes of emotions. To fulfil this task, she proposed a theory made of two phases, the first phase in which the stimulus is mapped into a core affect, and a second phase in which the core affect is categorised.

However, for Barrett, the categorisation of core affect is not a form of learning coming from experience, but something that helps to create and cause the emotive

experience in the subject ("core affect is differentiated by a conceptual act that is driven by embodied representations and available concepts"). The categories are not statics, but they depend on the perception of the emotive state on the basis of a previous conceptual knowledge. Furthermore, Barrett supports that the two phases are not sequential steps, but are two sources that influence each other until they reach a stable solution.

2.3 Psychophysiology of emotions: foundations

An emotional response consists of 3 different components: a behavioural component, a vegetative component and a hormonal component [43].

The *behavioural component* consists in appropriate muscular movements on the basis of the current stimulus.

The *vegetative component* facilitates the behavioural one, providing a rapid mobilization of the energy, to allow strong movements. For instance, the increasing of heart rate and changes of the diameter of blood vessel allows the blood to go to the muscles.

The *hormonal component* enhances the vegetative responses. Hormones secreted by the adrenal medulla (adrenalin and noradrenalin) augment the blood flow through the muscles and stimulate the conversion to glucose of the nutritive substances stored therein.

We summarized below a brief list of neural structures involved in affective processes [44].

Amygdala – The central nucleus of the amygdala is the most important region for the expression of emotional responses to harmful stimuli. For this reason is one of the most important structure in emotion research and it was proved that it has a crucial role in the processing of emotions: assigning an emotional value to the perceived objects and environment. After the destruction of the central nucleus, the animals do not show any signs of fear also if in presence of stimuli associated with harmful events. Vice versa, if the amygdala is electrically stimulated, the animals show signs of fear and a long stimulation causes stress diseases and gastric ulcer.

Orbitofrontal cortex – This area is known to be active in the recognition of emotion displayed by faces. It have also a role in the conditioning, in fact its activation is related to the value of an expected punishment or reward from a certain action.

Anterior cingulate cortex – The ACC is involved in the decision making and in premotor functions. It is important for its production of emotive responses, for example the arousal, as well as for the regulation of own emotive state and the perception of the pain.

Insula – This region is activated during the recognition and production of the disgust, however it is been discovered that it responds to sadness, fear and to re-

wards. It is involved also in aspects concerning sensation of pain.

Nucleus accumbens – This area, part of the ventral striatum, is involved in conditioning processes and in anticipation of rewards and punishments.

Thalamus – This structure transmits sensorial information to the rest of the brain.

Ventral tegmental area – In this area is generated a prediction of an error signal that is positive when a reward it is not expected and received, and negative when an expected reward is not received. This signal is controlled by a dopamine neurotransmitter.

2.4 Emotions in social interactions

Emotions are one of the most effective mediums of non verbal communication that humans have at their disposal to communicate simple but with big impact information. For instance a scream it is actually poor of direct information, nevertheless it gives to other subjects in the area indirect information of warning for an alert situation.

To understand better the communicative power of facial expressions, it is possible to do an experiment. Turn on the television on a channel with a film and remove the audio. Even without any dialog, it will be possible to understand what is currently happening by observing the facial expressions and gestures of the subjects in the video.

Several species of animals, including humans, communicate their emotions to the other through postural changes, facial expressions and non verbal sounds. These expressions allow to fulfil several social functions, for instance they communicate to the others what we are feeling and so what we are probably going to do [43].

An effective communication is a bidirectional process; this means that our expressiveness is useful only if others are able to collect our emotion cues and to interpret them. A study proposed by Kraut and Johnston [45] shows how humans are much more likely to smile when they are engaged in a social interaction with another person than when they are solitarily experiencing a pleasant emotion.

These early examples are sufficient to show the fundamental role of emotions in our social life. So, emotion researches are questioning if facial expressions configurations are innate or learned from environment. Darwin stated that emotional expressions are innate and to support this idea he collected a series of positive evidences. He observed the facial expressions of his sons and the expressions of member of other isolated cultures around the world. He argued that if people around the world display the same facial expressions of emotions, then these expressions have to be necessarily hereditary rather than to be learned, for the reason that a prolonged isolation of different communities of people leads to development of different languages, just because there are no biological basis for the development of the lan-

guage that justify the use of particular words for particular concepts. Conversely, facial expressions seem to be the same among the different cultures and this means that these are innate.

Ekman and his colleagues confirmed the idea of Darwin with others positive evidences observing the facial expressions of blind children, with respect to seeing children, and the emotive responses of isolated indigenous [46].

In addition to this innate behaviour of universal facial expressions after the input of a particular stimulus in the subject, it seems to exist also an imitative behaviour, in which mirror neurons play a crucial role.

Mirror neurons are activated when the animal do a particular task or when it is observing another animal which is doing the actions under consideration. This particular neural circuit is activated when we are observing another person that is doing a particular action and the feedback of this one could help us to understand what the other person is trying to do. For this reason it seems that mirror neurons are involved in the acquisition of the capacity of imitate other people's behaviour.

Therefore, according to some researchers, this biological component gives to us an *internal feedback* that helps us to understand what others are feeling when expressing an emotion through their faces and consequently it allows us to behave in a correct way during social interactions [43].

Imitation is probably one of the channels through which organisms communicate their emotions and regulate social interactions. For instance if we see someone looking sad, probably we will tend to display a sad facial expression too. The sensorial feedback contributes to put us in other's shoes and then it makes us more willing to provide comfort and support. This is probably one of the reasons of the pleasure of making people smile, because its smile makes smile us, rejoicing [43].

Emotions do not have only a role into communication processes, but it seems that they are crucial also for the evolution of society [19]. It is common to feel uncomfortable during the feeling of negative emotions, often not only psychics but also physical. Moreover, events that cause strong emotional responses are probably more remembered in the future [43].

Why do these negative emotions exist if they are so harmful to our mind and body? It seems that there is an analogy with the pain and the nervous system: if we touch a flame with a hand we burn ourselves and we feel pain, this is because in this way we are able to have an instant reaction and to remove the hand from the fire, limiting the damages to our skin and body. At the same way, emotions work like alarms to regulate the social life and allow the evolution of society [19].

For instance, if a person is reprimanded at work for not having well its duties, this one will feel embarrassment and then psychic and physical malaise. This in turn leads the subject to avoid this behaviour in the future, in order to not feel again this bad affective state. In this way, since the subject will work better from now on, the society obtains a possibility of evolution.

Conversely, positive emotions, lead the subject to feel pleasant feelings and then the research of these ones, thus leading welfare in society. For example, be of help to someone and then be reciprocated with a smile and thanks, leads us to a comfortable feeling, allowing us to repeat these behaviours in such a way to feeling again these sensations and creating a positive evolution for our society.

Summarizing, we have two fundamental roles of emotions: the first role is as external feedback that allows us to add to the message important features during the communication process, and the second role of internal feedback, which allows us to learn what kind of behaviours are advantageous and what are harmful, so that we are able to enhance our decisional process and the relationship with others during social interactions.

Chapter 3

A model for facial expression analysis

> A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding.
>
> William Gibson

3.1 Automatic affect recognition

For humans, affect recognition is a rather simple task. We are able to easily recognize an emotion through multimodal cues, such as the means of language, vocal intonation, facial expressions, hand gestures, head movements, body movements and postures [47]. It is not possible to say the same for today machines; in fact researchers are far from good results in this domain, where the reasons are due to several issues.

First of all we do not still have a unique and sound psychological model of emotions, and also neurosciences studies are not still able to lead us to a promising theory; thus, there are still present wide spaces of development without a shared and clear direction among the researchers.

Other issue involves the sensors currently available. Human perception is able to "produce" images and sounds of the world surrounding us with a high level quality, which is far from the quality of our electronic sensors. Often, computer vision researchers, have to face problems like the high dynamic range, image noise (especially in low illumination conditions), reflections and so on, causing a general decay of performance. Furthermore, we have still difficulties to develop techniques for objects recognition in images or natural language processing from audio, adding further barriers to the purpose achievement.

Finally, there is also a problem concerning the emotion datasets currently available. There are two possibilities to produce such datasets: using acted facial expressions or employing subjects engaged in real life behaviours. In the first case it is

possible to manage the scene and the subjects in order to produce videos of higher quality, however the expressions of the subjects are not genuine and often far from real life expressions: too much exaggerated or limited to a small and not enough informative set of expressions. In the second case it is possible to collect several genuine facial expressions, but the quality is not always good and data are too much heterogeneous for a regression procedure (see Chapter 5). Another problem of emotion datasets is the intrinsic difficulty to produce objective labels of emotions, for the reason that often are present a lot of different shades on emotion attribution, so several labels for the same facial expression are possible.

Early researches on emotions recognition dedicated more efforts to the recognition of the six basic emotions theorized by Darwin and afterward investigated by Ekman with cross-cultural studies [48]. The advantage of considering a well defined subset of categorical emotions is that the latter match well people's experience; in fact it is easy to categorize one of these prototype expressions in one of the basic category of emotions. However, reduce the recognition of emotions to a set of few categories it is not so interesting and informative for applications. Furthermore, these prototypical facial expressions consider only a small part of our every day social life, and become useless for applications that want to consider a human-computer social interaction.

For these reasons, researchers are changing direction by focussing their attention to the representation of emotion in a continuous space. This view derives from Russell's core affect theory that support the existence of a small set of (latent) variables able to describe all the possible emotions [41]. Researchers try to understand what kind of variables could be the best choice in order to describe emotions. Currently some examples of variables investigated by researchers are: valence, arousal, control, power, dominance. In particular valence and arousal are often used in several studies concerning emotions recognition, because they seem to reflect the main aspects of emotion [49]. The valence quantifies the positive or negative valence that the subject feels during an affect state. The arousal measures how the subject shows an active or a passive affective state.

By using these two variables it is possible to represent emotions as points in a 2D space, topologically divisible in four quadrants: the positive-active quadrant including emotions like joy or happiness, the negative-active quadrant comprehending emotions like anger or fear, the negative-passive quadrant enclosing emotions like depression and tiredness and finally the positive-passive quadrants having emotions like serenity and calm (Fig. 3.1).

Note that by considering such core affect variables as random variables of a continuous latent space, opens up the possibility of applying a variety of methods that in the last decade have been developed in statistical machine learning [50, 51, 52]. We will further discuss such option in later sections of the work presented here.

On the other hand, the problem of this approach is that if classification is specifically addressed, the task is not as intuitive as working in the discrete, categorical representation of six basic emotions. This issue produces a series of difficulties on the labelling task, which is necessary in order to have emotions datasets usable as training sets for the regression task.

Most of the works in the literature, either dealing with the classification into six

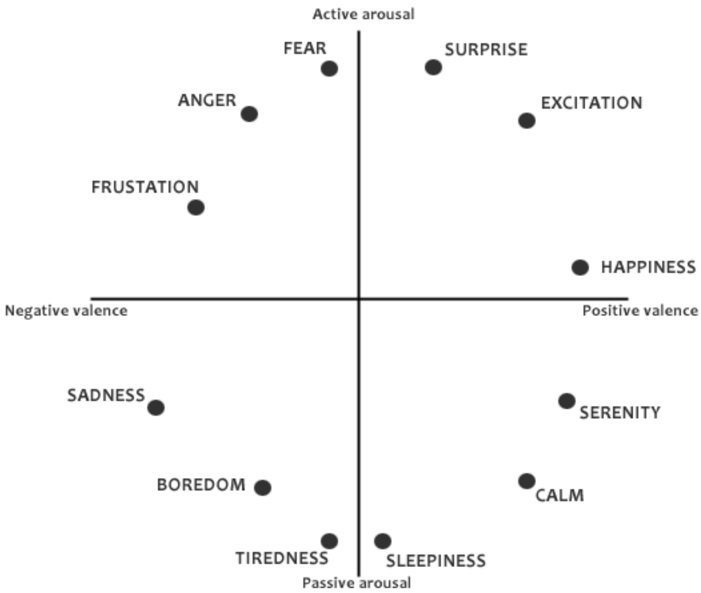


Fig. 3.1: Example of a 2D latent space of emotions

categories and the regression of a latent space of emotions, make use of visual signals for the recognition of the affect state. Several hypotheses are supported by psychologists and linguists about the importance of different cues in human affect judgment. Whereas it seems that the relative contributions of facial expression, speech and body gestures to emotion classification depend both on the current affective state of the subject and the surrounding environment, some studies support clues in favour to a major contribution of facial expressions in affect judgment [10]. Furthermore, the integration of multiple modalities, such as vocal cues, facial expressions and gestures, allows a better classification of emotions by humans [53].

It is possible to divide current research on facial expressions recognition in two main streams: the recognition of the affect and the recognition of facial Action Units (AUs). The facial AUs are descriptors of the movements of facial muscles [54]. When a subject produces a facial expression, this one can be described as a combination of a subset of AUs activation. As AUs are independent of interpretation, it is possible to use them as high-level decision-making process in order to recognize emotions. Ekman proposed a system to analyse AUs and map sets of AUs to particular affective states, the Facial Action Coding System (FACS) [54].

There are several methods for the classification of the facial features into categories of emotions or points of a latent space, many of these make use of well known regressors such as Support Vector Machine (SVM), Relevance Vector Machine (RVM), Neural Networks, Decision trees, and so on. Yet, there is no a rule for determining the type of regression to use for emotion recognition, all depends on the final purpose and expectations.

3.2 The proposed approach

Inspired by Russell's theory of core affect (see Section 2.2.6) our aim is to discover a mapping between the visible facial expression of a subject (activation of several facial muscles) and latent variables of a core affect space, so to describe the complex behaviour in a more simple way.

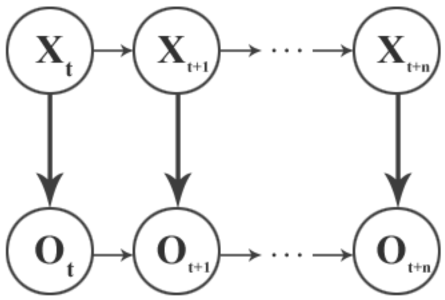


Fig. 3.2: Graphical model of the generative approach

We can consider our visual experience as a set of observations during the time, each one produce a latent variable of core affect[1]. An observation at time t is conditional dependent by observation at $t-1$. The same holds, in principle, for latent variables, namely a latent variable at time t is conditional dependent on latent variable at time $t-1$. If we consider this behaviour in a (probabilistic) generative framework, where the observation is "sampled" from the latent variable describing the latent state space, we can describe the process through a graphical model as depicted in Fig. 3.2.

Interestingly enough, the exploitation for recognition purposes of a generative model of visible expressions from an hidden core affect space shares some connections with the simulative approaches to social interaction: we are able to infer people affective states from their visible expressions, since we are ourselves capable to internally simulate "as if" (generate) such expressive behaviour, and compare simulated and actually observed behaviours with one another. This is for example one of the theories endorsed by mirror neurons theorists (cfr., Section 2.4).

In this work we are not interested to consider the affective valence of all the possible events and situations in our world, but only of facial expressions of subjects. For this reason we can interpret a frame of a video as an image with both a background content B_G, and a foreground content F_G. The foreground content contains the pixels useful for our purpose, namely the pixels of the face of the subject, whereas the background content includes all the other unnecessary pixels of the image. So our observation (the image) is conditioned by a foreground and a background random variables as in Fig. 3.3.

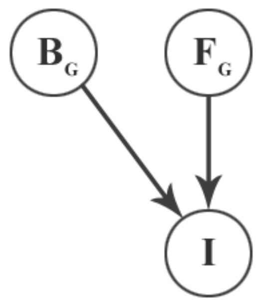


Fig. 3.3: Graphical model of an observed image I

If we consider the background as a constant, at most with some noise, it is possible to no longer consider it in the graphical model. Now extending our previously introduced model (Fig. 3.2) we obtain the general schema proposed in this work

[1] Remind that this concept is in general valid for every type of observation, in fact also a perception of an object or of a particular event has an affective valence, at most neutral

(Fig. 3.4).

The model outlined in Fig. 3.4 can be formalized as follows, by considering the time slice t, $t+1$. Let i_{t+1}, i_t, s_{t+1}, s_t, y_{t+1}, y_t, x_{t+1}, x_t denote samples from the random variables I_{t+1}, I_t, S_{t+1}, S_t, Y_{t+1}, Y_t, X_{t+1}, X_t, respectively

1. Sample the affective state at time $t+1$, conditioned on the previous affective state:
$$\widehat{x}_{t+1} \sim p(x_{t+1}|x_t). \tag{3.1}$$

2. Sample the feature based representation of the visual expression $\widehat{Y}_{t+1}$ on the basis of the current affective state and the visual expression at time t:
$$\widehat{y}_{t+1} \sim p(y_{t+1}|y_t, \widehat{x}_{t+1}). \tag{3.2}$$

3. Sample the state (position and scale) of the face inside the current frame, namely defining the region of interest (ROI), conditioned on the previous state of the face:
$$\widehat{s}_{t+1} \sim p(s_{t+1}|s_t). \tag{3.3}$$

4. Sample the observed scene (the frame at time $t+1$) from the current feature based representation of the visual expression, the current state of the face and the previous scene (we omit here for simplicity the constant background component B_g):
$$\widehat{i}_{t+1} \sim p(i_{t+1}|i_t, \widehat{s}_{t+1}, \widehat{y}_{t+1}). \tag{3.4}$$

The above sampling steps fully describe the probabilistic generative model of facial expressions from affective states. Clearly the actual aim of this thesis is to provide a method for inferring the hidden affective state x_{t+1} from the current feature based representation of the visual expression $\widehat{y}_{t+1}$ occurring in the observed scene $\widehat{i}_{t+1}$ in position defined by $\widehat{s}_{t+1}$.

In general terms such inference should be accomplished by "inverting the arrows", formally by using Bayes' rule:

$$p(x_{t+1}|i_{t+1}) = \frac{p(x_{t+1}, i_{t+1})}{p(i_{t+1})}, \tag{3.5}$$

where the two terms $p(x_{t+1}, i_{t+1})$ and $p(i_{t+1})$ may be obtained by marginalizing the joint probability:

$$\begin{gathered} p(i_{t+1}, i_t, s_{t+1}, s_t, y_{t+1}, y_t, x_{t+1}, x_t) = \\ p(i_{t+1}|i_t, s_{t+1}) p(s_{t+1}|s_t, y_{t+1}) p(y_{t+1}|y_t, x_{t+1}) p(x_{t+1}|x_t) \end{gathered} \tag{3.6}$$

in such a way that:

$$p(x_{t+1}, i_{t+1}) = \int p(i_{t+1}, i_t, s_{t+1}, s_t, y_{t+1}, y_t x_{t+1}, x_t) dI_t dS_{t+1} dS_t dY_{t+1} dY_t dX_t; \tag{3.7}$$

$$p(i_{t+1}) = \int p(x_{t+1}, i_{t+1}) dX_{t+1}. \tag{3.8}$$

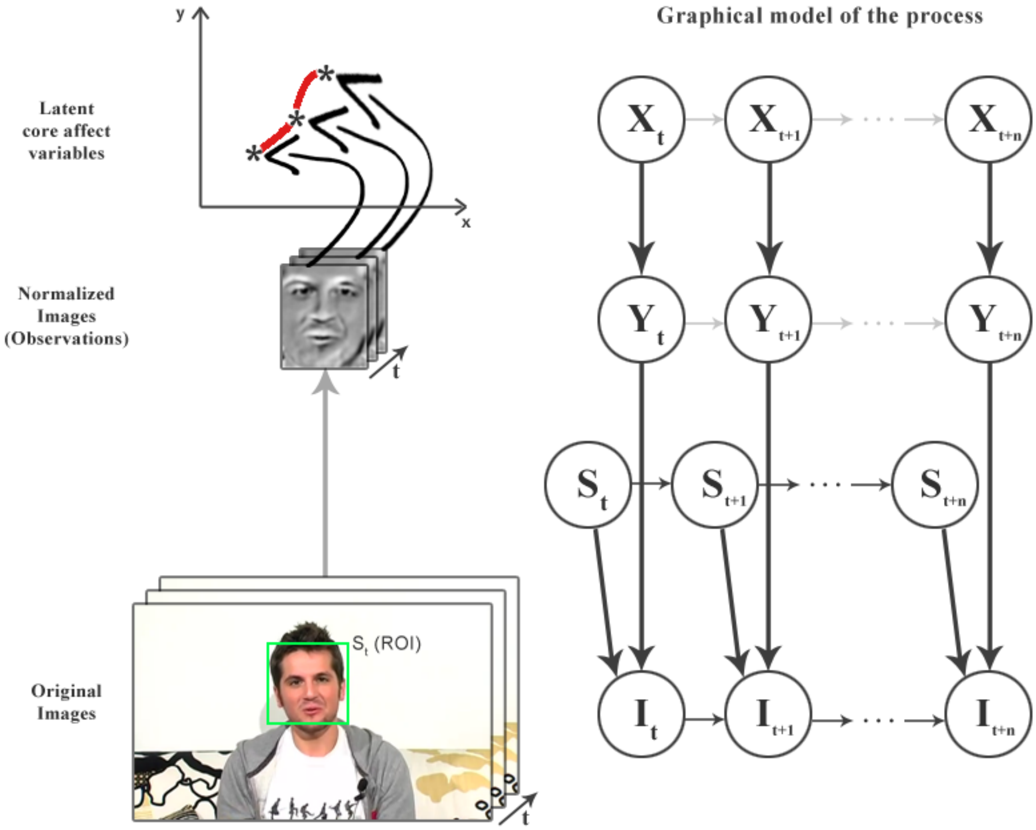


Fig. 3.4: Schema of the model proposed

Clearly, the integration of Eq. 3.7 and Eq. 3.8 cannot be computed in close form and some sort of approximation must be found.

The complexity of this inference/inversion task is not only formal but lies also in subtle technical issues. At the bottom of the process (cfr. Fig. 3.4) we have a frame of a video stream. It is clear the temporal dependence among consecutive frames, for the reason that we are considering a video stream with frames ordered over time. Obviously each frame of the video has a lot of useless parts, so it is necessary to extract only the region of interest (ROI) containing the face of the subject (the foreground variable presented in Fig. 3.3) defined by the current state of the face in the frame (s_t). However Eqs. 3.3 and 3.4 show that the ROI extraction must be accomplished over time. That means that ROIs inference must involve a filtering or tracking procedure. We will resort to Kalman filtering (see Section 4.3.1) to preserve the temporal conditioning between consecutive ROIs (Eq. 3.3).

Consider then the feature representation Y_{t+1} of the visible expression; we will assume in the simplest case that this is a normalized representation of the ROI (but we have experimented with more complex features, too). Indeed, the face of the subject is likely to be misaligned and/or suffer of different light conditions, so such normalization is necessary. Through this step we obtain a face picture of a fixed size and with eyes, nose and mouth placed in fixed positions. In this phase we miss the explicit temporal constraint between consecutive observations because, as it will be showed in the next chapter, whereas we make use of a Kalman filter to track facial landmarks positions (that are used to rotate and resize the face) over time, we do not have any temporal information for the light normalization task. Thus $p(Y_{t+1}|X_{t+1}) \simeq p(Y_{t+1}|Y_t, X_{t+1})$. However observing that ROIs are captured in a same video footage with approximately uniform light conditions, we can assume that these dependences are implicitely accounted for by lower level conditioning in the graph. In Fig. 3.4 we stressed this simplifying assumption by drawing the temporal dependences in lighter gray.

Finally a dimensionality reduction is made in order to describe the face expression with a few latent variables, and a core affect space is learned on the basis of a training set of facial expressions. Then we expect that an emotive episode is nothing more than a path along several points of this latent core affect space, so a classification can be made on the basis of the properties of these paths. Also in this step are no more considered temporal dependences between latent variables ($p(X_{t+1}) \simeq p(X_{t+1}|X_t)$), however with a regression process that maintains close points in the data space close also in latent space and vice versa, it is possible to infer an intrinsically temporal dependences between latent variables of consecutive frames.

In this chapter will be considered the process of dimensionality reduction, whereas in the next chapter will be treated the problem of face extraction and normalization, summarized with the name *"face sensing"*.

3.3 Desirable characteristics of the latent space

For our aim, the space of latent variables must have two characteristics:

1. Facial expressions that are similar must lie on close points of the latent space;

2. The transitions between neighbouring points in the latent space produce smooth transitions in the data space.

Furthermore the process of regression must to make use only of the face expression images without the use of labels, namely our observations are only the current pixels of the face.

The two first characteristics are necessary in order to set up a model where it is possible to easily investigate temporal dynamics of facial expressions. If similar facial expressions lie on similar location, it is possible to expect that the path created by a set of consecutive frames generates a smooth and easy to model spline.

Modelling each temporal facial expression with a spline, allows to remove particularly difficult issues related to the duration of an affective state, that is one of the major challenging in emotion recognition [10]. Furthermore, as our work want to investigate affective states in a social view, it is possible to check the existence of a model able to describe interactions between splines, which can be used to recognize and predict the affective character of a social interaction occuring between two subjects.

Finally, using unlabelled information it is crucial in order to simplify the data collection task. By creating a mapping directly between feature points and latent variables (variables not chosen by other people and so not biased) it is not necessary to label the audio-visual material that often is a quite difficult task.

First of all the labelling task need the decision of the latent variables to use during the evaluation process, this implies the need to choose these variables and not other, unfortunately we have already told that at the moment there is not a unique and clear psychological theory on that issue.

The second problem is that the labelling process, even if it is done with the use of several subjects, it is a subjective task and the use of graduated scales are not sufficiently precise for producing good quality training data. In our vision, as we told previously, the classification task, namely the labelling process, has to be done in a next phase investigating the regressed latent space of emotions (a similar approach is proposed by Huang et al. [55]).

This last point suggests a procedure of unsupervised learning, where, as we told before, a dimensionality reduction algorithm can be the right approach. However it is important that during the process of dimensionality reduction the variables of the latent space remain sufficiently informative to reconstruct the observations, and that the topology of the new latent space produces smooth dynamics among all the facial expressions. For these reasons linear dimensionality reduction techniques are not sufficient to our purpose.

At a first sight it may seem necessary to extract only a few representative features from the raster image of the face in order to reduce the dimension of the observations and consequently having more chances to regress a performant mapping from data space to latent space, however doing this way we only miss important information of the face texture, such as facial wrinkles. Furthermore, techniques such as Active Shape Models or Active Appearance Models (see Section 4.4) introduce a further noise on the observations, being the facial features extraction a not so easy task. So, the use of the pixel values of the face image may be an appropriate choice for the task of complex facial expressions recognition.

In this work we propose to use Gaussian processes and in particular of Gaussian Process Latent Variable Model (GPLVM) as method of dimensionality reduction, since they possess several interesting properties useful to achieve our aim, as will be clearer in next sections.

3.4 Background on Gaussian processes

Gaussian processes (GP) are hugely powerful tools for regression, nevertheless they play an important role in the theory of probability. Gaussian processes provide a principled, practical, probabilistic approach to learning in kernel machines [50]. They are used in several domains, for example one application domain is geostatistics, a branch of statistics studies phenomena with spatiotemporal character. Another application of GP is to model things evolving over time, for example a face expression given a set of video frames, like what we are investigating in this work, or the 3D pose of a person given a series of 2D silhouette [56]. Let us now formally define a Gaussian process:

Definition 1. *For any set X, a Gaussian Process on X is a set of random variables $(f(x) : x \in X)$ s.t. $\forall n \in \mathbb{N}$ and $\forall x_1, ..., x_n \in X$, $(f(x_1), ..., f(x_n))$ is a multivariated Gaussian distribution*

Whereas a probability distribution describes the distributions of scalars and vectors, a *stochastic process* describes distribution of functions. Simplifying, it is possible to think a function $f(x)$ as a very long (infinite) vector where each entry in the vector is the value of the function at x. A Gaussian process defines a *prior over functions*; in fact it is a generalization of Gaussian distributions and, as a stochastic process, it describes a distribution of functions. To clarify better how GP works in practice, we will now introduce a trivial example:

Example 1. *Let $X = \mathbb{R}$ and $W \sim \mathcal{N}(0, 1)$, than $f(x) = xW$ is a Gaussian process*

In this example we have realized a set of random lines (Fig. 3.5). We have defined a model of function ($f(x)$) which is modulated by a normal distribution on W, generating a set of normal distributed linear functions.

As a Gaussian distribution over functions, a Gaussian process can be uniquely described by its mean and covariance *functions*, respectively $m(x)$ and $k(x, x')$ defined as:

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)], \\ k(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))], \end{aligned} \tag{3.9}$$

that is a real process $f(x)$ can be defined as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{3.10}$$

if and only if for all $n \geq 1$ and $x_1, ..., x_n$, $(f(x_1), ..., f(x_n)) \sim \mathcal{N}_n(\mu, K)$, with $\mu = [m(x_1), ..., m(x_n)]$ and $[K]_{ij} = k(x_i, x_j)$.

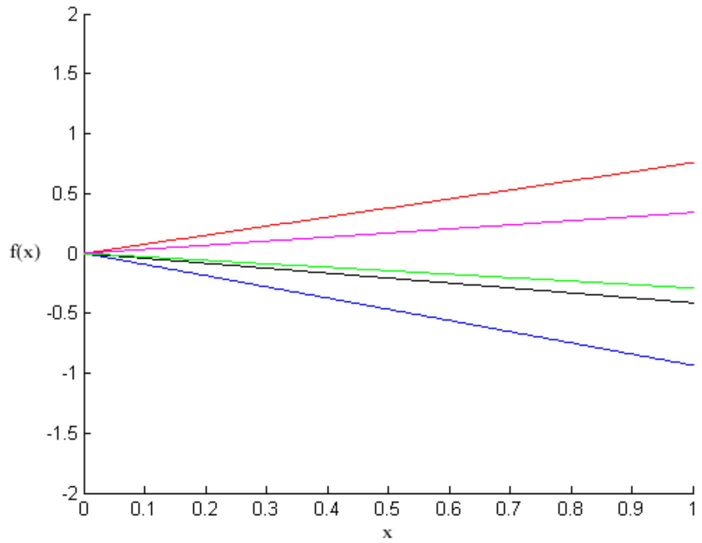


Fig. 3.5: Example of random lines drawn from GP prior

Usually the mean function is specified as a zero function, whereas the covariance function is often a non-linear one. It is clear that the covariance function plays an important role in the regression task, since it encodes all the necessary information on $f(x)$, such as its local smoothness, continuity, periodicity, etc.

If we specify a zero function as the mean function of the Gaussian process, the covariance $k(x, x')$ results:

$$k(x, x') = \mathbb{E}[f(x)f(x')] \tag{3.11}$$

that is the covariance function is completely defined through dot products (always thinking $f(x)$ as a long vector with infinite dimension). Here, if there is a *kernel function* K such that $K(x_i, x_j) = f(x_i)f(x_j)$, it is possible to use only K in the training algorithm, without even esplicity know what f is. A kernel function must be continuous, symmetric, and positive semi-definite Gram matrix.

The trivial case is a *linear kernel*, where $K(x_i, x_j) = x_i^T x_j + c$ with c a constant. With this kind of kernel, as covariance function of a GP, it is possible to describe a set of linear functions, such as those in the example 1. However this model is too rigid for most of the regression tasks, for the reason that let us to regress only linear functions.

A Gaussian process prior (Eq. 3.10) allows for *non-linear* mappings if the kernel k is non-linear. There are several kind of non-linear kernels with several free parameters in order to describe sufficiently flexible models for regression. An example of non-linear kernel is the *Radial Basis Function* (RBF) defined as:

$$k(x_i, x_j) = \sigma_s^2 \exp(-\frac{1}{2\ell^2}(x - x')^2) + \sigma_n^2 \delta_{ij} \tag{3.12}$$

with the length-scale ℓ, the signal variance σ_s and the noise variance σ_n as free parameters of k. These free parameters control the properties of the functions generated by the GP and are called *hyperparameters*.

In a simple task of regression it is possible to assume a fixed set of values of the free parameters. In this way we are restricting the regression to only those functions with specific properties given by the selected set of hyperparameter's values, and consequently we are defining a bias for the regression; this is what happens with other machine learning methods such as Support Vector Machine or Neural Networks, where in fact the most difficult problem is to define a good set of parameters for the model in order to allow the algorithm to learn the objective function.

As it will be shown further, with GPs it is possible to regress also the best set of hyperparameter's values that fit better the observed data, augmenting the flexibility and consequently the quality of the overall regression. However, now we start to define a simple regression task over a GP assuming the absence of noise on the observations.

Definition 2. *Given a set $X_\star$ of input points and a kernel function $K(X, X')$ we define a GP $f_\star \sim \mathcal{N}(0, K(X_\star, X_\star))$ representing our prior. Let f be a set of observation on a subset $X \in X_\star$. Then, the joint distribution of the observations set f and the prior $f_\star$ is:*

$$\begin{bmatrix} f \\ f_\star \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} K(X, X) & K(X, X_\star) \\ K(X_\star, X) & K(X_\star, X_\star) \end{bmatrix}) \tag{3.13}$$

and consequently the regression task is defined as the computation of the posterior:

$$f_\star | X_\star, X, f \sim \mathcal{N}(m_{post}, k_{post}) \tag{3.14}$$

with m_{post} and k_{post} computed as:

$$\begin{aligned} m_{post} &= K(X_\star, X)K(X, X)^{-1}f \\ k_{post} &= K(X_\star, X_\star) - K(X_\star, X)K(X, X)^{-1}K(X, X_\star) \end{aligned} \tag{3.15}$$

If our observations are corrupted by noise it is sufficient to consider our observation as $y = f(x) + \epsilon$ with ϵ an additive independent identically distributed Gaussian noise with variance σ_n^2. For further details on this less trivial approach we remand to [50].

As it was told previously, it is possible to regress not only the parameters w of the kernel function, but also its hyperparameters θ. This task is usually called *model selection*, where the model is usually indicated by $\mathcal{H}_i$.

Also in this case, it is possible to use a Bayesian approach applying a series of Bayes' rules in a hierarchical way, namely inference takes place one level at a time. At the bottom level we have to compute the *posterior over the parameters* as:

$$p(w|y, X, \theta, \mathcal{H}_i) = \frac{p(y|X, w, \mathcal{H}_i)p(w|\theta, \mathcal{H}_i)}{p(y|X, \theta, \mathcal{H}_i)} \tag{3.16}$$

where $p(y|X, w, \mathcal{H}_i)$ is the likelihood and $p(w|\theta, \mathcal{H}_i)$ the prior, encoding a probability distribution of our knowledge about the parameters prior. The normalizing constant $p(y|X, \theta, \mathcal{H}_i)$ is independent of the parameters and called the marginal likelihood, defined as:

$$p(y|X, \theta, \mathcal{H}_i) = \int p(y|X, w, \mathcal{H}_i)p(w|\theta, \mathcal{H}_i)dw \tag{3.17}$$

A level above we can similarly express the *posterior over the hyperparameters* as:

$$p(\theta|y, X, \mathcal{H}_i) = \frac{p(y|X, \theta, \mathcal{H}_i)p(\theta|\mathcal{H}_i)}{p(y|\mathcal{H}_i, X)} \tag{3.18}$$

where it is possible to see as current likelihood the marginal likelihood from the previous level, whereas $p(\theta|\mathcal{H}_i)$ is the hyper-prior, that is the prior for the hyperparameters, and $p(y|\mathcal{H}_i, X)$ is a normalizing constant defined as:

$$p(y|\mathcal{H}_i, X) = \int p(y|X, \theta, \mathcal{H}_i)p(\theta|\mathcal{H}_i)d\theta \tag{3.19}$$

Finally, at the top level we are able to compute the posterior for the model as:

$$p(\mathcal{H}_i|y, X) = \frac{p(y|X, \mathcal{H}_i)p(\mathcal{H}_i)}{p(y|X)} \tag{3.20}$$

where $p(y|X)$ is now simply defined as:

$$p(y|X) = \sum_i p(y|X, \mathcal{H}_i)p(\mathcal{H}_i). \tag{3.21}$$

Gaussian processes place a prior on the space of functions f directly, without parameterizing f. Therefore, Gaussian processes are non-parametric.

3.5 Gaussian Process Latent Variables Model

Latent Variable Models (LVMs) carry out the idea that data which is apparently high-dimensional may actually lie on a low-dimensional non-linear manifold. Considering a set of observation $(y_1, ..., y_n) \in \mathcal{Y}^D$ and a set of latent hidden variables $(x_1, ..., x_n) \in \mathcal{X}^L$ with $L \ll D$, we wish to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ such that $\forall i \in \{1, ..., n\}, y_i = f(x_i, W) + \epsilon$ as W free parameters of f and ϵ an additive noise over the observations.

From what was discussed in the previous section it is clear that GPs are powerful tools for regression for several reasons, one above all the possibility to regress a model without parameterizing the prior. Furthermore, their ability to works with several kinds of kernel functions permits to specify the desired properties of the target function.

It would be interesting to use GPs as a LVM in order to obtain on the one hand the dimensionality reduction and on the other hand specific properties over the latent space created by the dimensionality reduction process, as well as the possibility to solve the problem in a sound probabilistic way.

A probabilistic interpretation of the problem allows for example to handle incomplete data, to sample new data from the probabilistic model learned and to extend the model with prior knowledge or integrate it with other probabilistic models.

Lawrence goes in this direction and propose a new dimensionality reduction approach based on Gaussian Processes: the Gaussian Process Latent Variable Model (GP-LVM) [57].

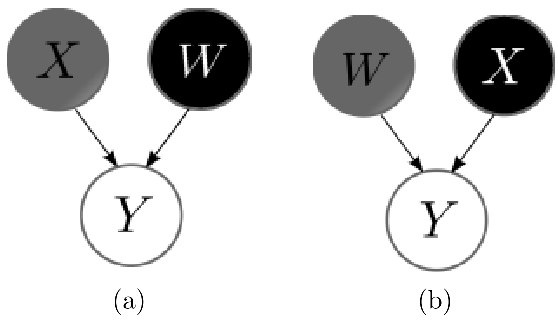


Fig. 3.6: Graphic model of PPCA (a) and GPLVM (b)

Lawrence shows that Principal component analysis (PCA) and Probabilistic PCA (PPCA) are nothing more that particular cases of GP prior with linear kernel. Using a GP prior with a non-linear kernel it is possible to face the problem of dimensionality reduction with non-linear mappings.

Whereas the PPCA combines a Gaussian likelihood

$$p(Y|W, X, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|Wx_n, \beta^{-1}I), \tag{3.22}$$

with a Gaussian prior on the latent variables X marginalising over it:

$$p(y_n|W, \beta) = \int p(y_n|x_n, W, \beta)p(x_n)dx, \tag{3.23}$$

GPLVM does the opposite (Fig. 3.6) placing the prior on the parameters W of the mapping function $p(W) = \prod_{i=1}^{D} \mathcal{N}(w_1|0, \alpha^{-1}I)$ and marginalising over it:

$$p(y_n|X, \beta) = \int p(y_n|x_n, W, \beta)p(W)dW \tag{3.24}$$

where consequently the solution for X can be found by assuming that y_n is i.i.d and maximising the likelihood

$$p(Y|X, \beta) = \prod_{n=1}^{N} p(y_n|X, \beta) \tag{3.25}$$

and finally obtain a marginalised likelihood for Y:

$$\begin{aligned} p(Y|X, \beta) &= \prod_{n=1}^{N} \int p(y_n|x_n, W, \beta)p(W)dW \\ &= \frac{1}{(2\pi)^{\frac{DN}{2}} |K|^{\frac{D}{2}}} \exp(-\tfrac{1}{2}Y^T K^{-1} Y) \end{aligned} \tag{3.26}$$

where $K = \alpha XX^T + \beta I$ and $X = [x_1^T ... X_n^T]$. Maximising Eq. 3.26 is equivalent to minimising its negative logarithm

$$L = -\frac{DN}{2} \ln(2\pi) - \frac{D}{2} \ln |K| - \frac{1}{2} tr(K^{-1}YY^T). \tag{3.27}$$

It is possible to optimise the likelihood with respect to X with the gradient

$$\frac{\partial L}{\partial X} = \alpha K^{-1} Y Y^T K^{-1} X - \alpha D K^{-1} X, \tag{3.28}$$

which implies the solution

$$\frac{1}{D} Y Y^T K^{-1} X = X \tag{3.29}$$

and then with some algebric manipulation of this formula leads to

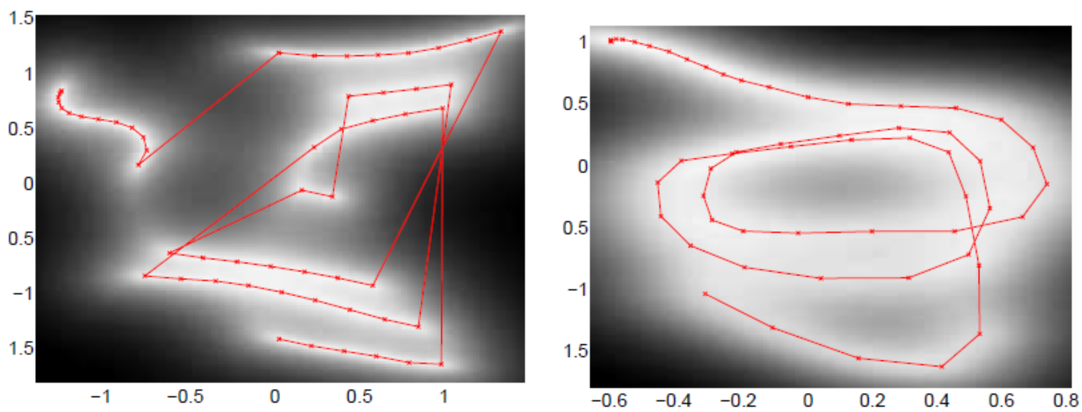
$$X = U_q L V^T \tag{3.30}$$

where U_q is a $N \times q$ matrix whose columns are eigenvectors of YY^T, L is a $q \times q$ diagonal matrix whose jth element is $l_j = (\frac{\lambda j}{\alpha D} - \frac{1}{\beta \alpha})^{-\frac{1}{2}}$ and V is an arbitrary $q \times q$ orthogonal matrix. This eigenvalue problem can easily be shown to be equivalent to that solved in PCA, for this reason PCA inner products YY^T can be replace by a non-linear kernel in order to extend PCA model with non-linear mapping using the same approach shown above.

Lawrence suggests an RBF function (Eq. 3.12) as non-linear kernel and then the use of a scaled conjugate gradients (SCG) [58] for the task of non-linear optimisation.

Furthermore, in order to make the algorithm computationally more efficient was used a *sparsification process*, sampling data points using informative vector machine (IVM) [59], which subsamples the observations sequentially according to the reduction in the posterior process' entropy that they induce.

GP-LVM gives a smooth mapping from latent to data space, however, whereas points that are close in latent space will be close in data space, points close in the data space *may not* be close in latent space. This is due to the intrinsically non-linear nature of the mapping that cannot maintain the correct distances between points.



(a) (b)

Fig. 3.7: Examples of the application of back-constraint. The figure (a) show a latent space regressed without the use of back-constraint, whereas the figure (b) show the same latent space regressed with back-constraint. In the last case it is clear a smoother dynamics of the points.

To solve this issue Lawrence and Candela [60] propose a way to force GP-LVM in order to fulfil the second property. By back constraining each x_i to be a smooth mapping from y_i local distances can be respected. As mapping it is possible to use any smooth function such as Neural network, RBF Network or Kernel based mapping. The constraints are of the form

$$x_i = g(y_i, W)$$

with W parameters of the smooth mapping function. This constrains points that are close in the observed space to also be close in the latent space. The mapping from Y to X is called *back-constraint*.

It is then possible to extend GP-LVM with the back-constraint computing the gradient $\frac{\partial L}{\partial W}$, with L negative log likelihood of GP-LVM, via chain rule and optimise parameters of back-constraining mapping (Fig. 3.7).

Chapter 4

Face sensing

> A man's face as a rule says more, and more interesting things, than his mouth, for it is a compendium of everything his mouth will ever say, in that it is the monogram of all this man's thoughts and aspirations.
>
> Arthur Schopenhauer

4.1 Introduction to the architecture

Crucial element of this project is the detection and extraction of normalized faces from videos, namely *face sensing*. Without face pictures it is impossible to train an emotions recognition system, and obviously also the performance of the face extractor are an important factor to the success of the project.

The architecture of the face extractor includes four main blocks (Fig. 4.1):

Face detector – The first step for the extraction of facial cues from videos is to detect a face in it. Fortunately, the current state of the art let us to obtain good results with low computational costs, especially in good illumination conditions of the scene.

Tracking system – This element is necessary in order to filter the errors of the measurement system (face detector) or to reconstruct missing measures with predictions. With a tracking system we are able to further improve the performance of face detection in a video sequence.

Facial landmarks localization – It is necessary to have some reference points in order to recognize the pose of a face and to consequently extract a fixed and normalized area containing this one. Fortunately the human face has eyes, mouth and nose which can be easily used as reference points.

Facial image normalization – Using the information of the facial landmark detector it is possible to compute a rigid transformation to align the eye line with

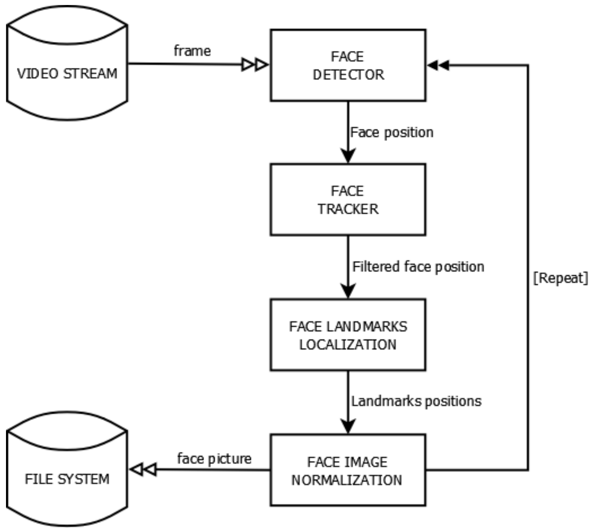


Fig. 4.1: Architecture of the face extractor system

the X-axis of the image and scale it in order to obtain fixed position and size of the face elements. Furthermore, it is necessary the application of an algorithm able to normalize the images under different lighting conditions. In fact we assume that in a video shot the lighting condition are acceptable and more or less stable (see below), but this assumption does not consider the problem of different lighting conditions among all the videos in the database.

In uncontrolled environments there are several variables that are difficult to manage, for this reason some assumptions were made in order to simplify the problem:

1. In a video shot is present only one subject. With this assumption it is possible to avoid the problem of identity measurement, which is necessary if we want to track several persons at the same time. Furthermore we assume that the subject is always present in the scene from the beginning of tracking;

2. The subjects do not present beard or occlusions due to eye glasses, hair, hats, etc. This assumption lets us to have more precision during the process of pose estimation and affect recognition;

3. The distance of the face from the camera remains more or less at the same distance for the entire duration of the video shot. Furthermore, the head remains most of the time in a frontal position. With this assumption it is possible to extract only faces of a sufficient scale dimension and only frontal, so that it is possible to obtain good facial cues for emotion recognition;

4. The illumination of the scene is sufficient to easily detect a face, that means the restraint of shadows on the face and/or highlights. With this last assumption the performance of the face detector can be acceptable without the use of pre-processing methods for images enhancement.

It is clear that these assumptions make us away from real life problems, however

addressing these issues now it would be premature and could be done in future works when the model for emotions recognition will be sufficient sound.

All the components of the architecture have been programmed in C++ using OpenCV[1] libraries. In the following sections each component will be discussed and analysed.

4.2 The face detector

As has been said in the previous section, face detection is the first step in order to extract faces from video streams. Its performance have a decisive impact on overall efficacy. An optimal face detector should be able to locate all the faces present in an image regardless of different scale, orientation, pose, expression and so on. It is obvious that we are far from this optimal behaviour, however with the current state of the art we are able to obtain good results, especially in simplified and well known scene conditions, as in this work happens.

Face detection can be accomplished in several ways. The easiest but less efficacious way is using the information of skin colour [61, 62, 63]. These methods works well when the background colour of the scene is well separated by the colour of the subject's skin and when in the scene are not present objects with a colour similar to that of the skin, otherwise performance decrease drastically. Also lighting conditions can affect the results. On the contrary, these methods are able to consider different orientation, pose and size of faces without additional effort.

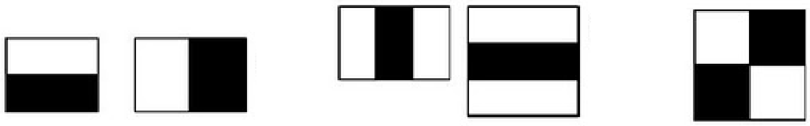
Other methods consider the motion as a cue to extract faces from videos, for the reason that faces are usually moving objects [64]. These methods consider several frames to detect moving entity. However, faces are not the only type of objects that are able to move in a video stream; therefore these kinds of detectors must use also other approaches to discriminate among moving entities, otherwise they could fail the detection. A method to detect faces and not other moving objects is to detect a blinking pattern of the eyes to exclude other type of moving entities [65].

The last class of methods uses facial shape or facial appearance as cue for face detection. The input image is scanned at all possible locations and scales by a sub-window, then a trained face detector is used in order to classified the pattern inside the sub-window as face or non-face. In the works of Viola and Jones [66], techniques like integral image and training of the face detector using AdaBoost-based methods allow to further improve speed and accuracy of the detection, becoming the state of the art for face detection.

4.2.1 Viola and Jones' face detector

Viola and Jones' face detector classifies images based on the value of simple features, namely Haar basis functions. In details, these features are classified in three classes: *two-rectangle features*, *three-rectangle features* and *four-rectangle features*. The value of a two-rectangle feature is the difference between the sum of the pixels values within two rectangular regions. The value of a three-rectangle feature is the difference

[1]http://opencv.org/



(a) (b) (c)

Fig. 4.2: Examples of Haar features. From left to right: Two-rectangle features (a), Three-rectangle features (b), Four-rectangle feature (c).

between the sum within two outside rectangles subtracted from the sum in a centre rectangle. Finally a four-rectangle feature computes the differences between diagonal pairs of rectangles (Fig. 4.2). The regions of a feature must have the same size and shape and be horizontally or vertically adjacent.

Viola and Jones introduce a new image representation called *integral image* that allows a fast evaluation of the Haar features during the detection process. The integral image at location x,y contains the sum of the pixels above and to the left of x,y inclusive.

With the integral image previously computed, it is possible to calculate any rectangular sum using four reference points of the integral image itself, making the computation process very fast. For each considered sub-window there can be thousands of this kind of feature, which can be seen as mappings from a space $N \times N$, the dimension of the sub-window, to scalars $z_k(x) \in \mathbb{R}$. These scalar numbers create an overcomplete feature set that can be used to train the system.

The training process is taken by an AdaBoost learning procedure [67] that has the aim of learning the best sequence of weak classifiers $h_m(x)$ and the relative combining weights α_m in order to obtain a strong classifier H_M defined as:

$$H_M(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \tag{4.1}$$

So, the AdaBoost algorithm is used to solve three fundamental problems: (1) selecting effective features from a large feature set; (2) constructing weak classifiers, each of which is based on one of the selected features; and (3) boosting the weak classifiers to construct a strong classifier [68].

For each example into training set is associated a weight w_i. This set of weights represents the distribution of the training examples. After each iteration, the training examples that results harder to classify are given larger weights w_i in order to give more importance to these examples at next iteration. For further details about the AdaBoost algorithm see [67].

To further improve the accuracy of the face detector, Viola and Jones propose a trained cascade of strong classifiers instead of a one single trained strong classifier as a solution for reducing false alarm rate. The idea is to train a cascade of strong classifiers in this way: the first strong classifier is trained with all the positive and negative training examples, then the next strong classifiers are trained using non-face examples that pass though the previously trained cascade.

When a sub-window has to be passed to the cascade in order to determine if it contains a face pattern or not, each cascade is questioned, then the features pass

though the next cascade node only if the previous node classify them as a face. If the sub-window is classified by all the cascade nodes as a face, the overall answer of the face detector will be positive.

The power of this approach is that can be used not only for face detection, but also for other kind of objects detection. For this purpose the only things needed are a new training set with example of the object to detect and a new set of Haar features that can be effective for the detection task. For this reason, for example, it is possible to create several cascade file each one specialized in a particular pose of the face, such as frontal and profile faces. This is the approach of our work; several cascade files are used in order to detect a face even if it is not frontal, so that it is possible to miss a very few measurements, representing crucial information for the tracking systems in order to improve the overall performance of the face extractor system.

4.3 The tracking system

In order to improve the accuracy of the face detector and to remove noise due to missing measures (e.g. occlusions) or simply due to face detector failures, a tracking system can be used.

A tracking algorithm has the aim of localizing a moving object inside a video stream. It is composed of two main steps: a first step of *prediction* and a second step of *correction*.

In the first step the tracking system make a prediction of where the considered object could be. This prediction is made in a recursive manner taking in consideration the previous predictions and can be seen as a probability distribution with its mean and covariance, so usually it is used the mean as value of the location of the object.

In the second step a correction to the previous prediction is made. This correction takes into account the quality of the current measure and updates the current probability distribution of the states.

There are several algorithms for video tracking, however the most used in computer vision are based on Kalman filter [69] or Particle filter (known also as CONDENSATION algorithm [70]).

The Kalman filter and Particle filter are based on similar ideas and probabilistic models, however Kalman filter works in an optimal way only for dynamic linear models and it assume that all error terms and measurements have a Gaussian distribution, whereas Particle filter generalize the model in order to capture also non-linear dynamics and non-Gaussian distributions.

For this reason the Particle filter is more complete, however it is also computational more expensive, so a Kalman filter is preferable if the necessary conditions subsist.

In the case of face tracking these conditions can be valid only simplifying the problem with further assumptions. The first assumption is that the errors and the measurements must be Gaussians and the second assumption is that the subject in the video does not compute extreme movements with its head or occlusion of the

face.

We can easily assume that for each frame the position of the subject (state) can be represented by a Gaussian distribution, where the mean indicates the current location of it. Also the errors can be interpreted as Gaussian noise, so that the first assumption is solved.

The problem is on the second assumption; in fact it is possible to assume a linear dynamics of the subject's face only in controlled conditions. In real life a subject can for example sneeze or compute a rapid and non-linear movements with its head, creating serious problem for a Kalman filter based tracker. Not only rapid non-linear movements can reduce accuracy of a Kalman filter, but also occlusions of the face that in real life are more common (consider for example the occlusion created by the hands on the face when the subject is stressed or tired).

In this work are considered relatively simple videos with only one subject computing simple movements generalizable with linear dynamics models. Sometimes occlusions of the face happen, but are sporadic. For this reason a Kalman filter based tracker can be a good solution for our purpose, however in order to face possible future works where the two assumptions fails also a Particle filter based tracker was considered and then an evaluation of accuracy of the two tracker was made.

In the next two paragraphs will be introduced the theoretical fundamentals of Kalman and Particle filter and the relative tracking systems used in the project. Then a further paragraph will show the results of a test for the accuracy of the two tracking systems.

4.3.1 Kalman filter based tracker

The Kalman filter has the aim to estimate the state $s \in \mathbb{R}^n$ of a discrete time process governed by the equation:

$$s_t = As_{t-1} + Bu_{t-1} + w_{t-1} \tag{4.2}$$

where

A is the transition matrix of the model and it is applied to the previous state s_{t-1}. This matrix correlates the state at time $t-1$ and the state at time t and it is responsible of the state update.

B is the control matrix on the input of the system. It is applied to the control vector $u_{t-1} \in \mathbb{R}^k$ and maps it to the dimension of the state s.

$w_t \in \mathbb{R}^n$ is the matrix of the noise on the state. It is assumed as a Gaussian with mean 0 and covariance described by the matrix Q_t.

At time t an observation of the real state s_t is made through the measure vector $z \in \mathbb{R}^m$ that is modeled by:

$$z_t = Hs_t + v_t \tag{4.3}$$

where

H is the matrix mapping the measures space to the state space, that is the matrix H describe how the measure is created starting from the state.

$v_t \in \mathbb{R}^m$ is the noise of the observation and it is described by a Gaussian with mean zero and covariance described by the matrix R_t.

Kalman filter is a recursive estimator where the phases of prediction and correction occur in a cyclic way. The predict phase projects the state and the covariance of the state from time $t-1$ to time t. The equations for this step are:

$$\tilde{s}_t = As_{t-1} + Bu_{t-1} \tag{4.4}$$

$$\hat{P}_t = AP_{t-1}A^T + Q_{t-1} \tag{4.5}$$

Where P is a matrix describing the covariance of the error on the estimation of the state. In the correction phase the a priori estimation of the state obtained in the predict phase and the new observation are used in order to correct the prediction. The equations for this step are:

$$K_t = \hat{P}_t H_t^T (H_t \hat{P}_t H_t^T + R_t)^{-1} \tag{4.6}$$

$$s_t = A\tilde{s}_t + K_t(z_t - H_t A\tilde{s}_t) \tag{4.7}$$

$$P_t = (I - K_t)\hat{P}_t(I - K_t)^T + K_t R_t K_t^T \tag{4.8}$$

The entire process can be summarized as:

1. Project the state ahead (eq. 4.4);
2. Project the error covariance ahead (eq. 4.5);
3. Compute the Kalman gain (eq. 4.6);
4. Update the estimation with measurement z_t (eq. 4.7);
5. Update the error covariance (eq. 4.8).

In this work the state is described by the x,y position of the face into the frame of the video and the size w of the square window containing the face.

Assuming a constant speed for each of the 3 component of the state and the absence of control input, it is possible to describe the vector s and the matrix A in this way:

$$s = \begin{bmatrix} x \\ y \\ w \\ dx \\ dy \\ dw \end{bmatrix}, A = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

As measure system it is used the same Viola and Jones' face detector. This may seem counterintuitive and the same thing of detecting the face frame by frame, however

it is important to remember that in this phase we are modelling a filter and that the measures are affected by noise and that sometimes are missing, so the Kalman filter allow us to correct these measures (and to reconstruct the more probable state) in order to have better performance. Consequently, the measure vector z and the transformation matrix H are described as:

$$z = \begin{bmatrix} x \\ y \\ w \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

4.3.2 Particle filter based tracker

Particle filter, like Kalman filter, has the aim of estimating the state of a movable object, but unlike Kalman filter, this algorithm results optimal on the tracking of non-linear trajectories or dynamics affected by non-Gaussian noise.

Also this algorithm relies on a two steps procedure of prediction and updating. These steps derive from Bayes rule which describe the filtering distribution though the likelihood $p(z_t|s_t)$ and predictive density $p(s_t|z_{1:t-1})$, so that:

$$p(s_t|z_{1:t}) \propto p(z_t|s_t)p(s_t|z_{1:t-1}) \tag{4.9}$$

Usually the likelihood function is known, whereas the predictive density is not, for the reason that it is given as an hard to compute integral:

$$p(s_t|z_{1:t-1}) = \int p(s_t|s_{t-1})p(s_{t-1}|z_{1:t-1})ds_{t-1} \tag{4.10}$$

Analytical solutions are available only in a few simplified cases, for example when the model is linear and Gaussian (this is the example of Kalman filter). However, in general $p(s_{t-1}|z_{1:t-1})$ is a complicated function and simulation methods, like Monte Carlo methods, are required.

The method uses a sample-based approach to estimate the probabilistic distribution of the state (Monte Carlo method). The posterior probability is represented by a set of randomly chosen weighted samples, so that:

$$p(s_t|z_{1:t}) \approx \sum_{i=1}^{N_s} w_t^i \delta(s_{0:t} - s_{0:t}^i)$$

and

$$\sum_{i=1}^{N_s} w_i = 1$$

Where δ is a Dirac function and $\{w_t^i, s_t^i\}_{i=1}^{N_s}$ denote a set of weights and particles. It is then possible to approximate the integral in equation 4.10 with a sum of a discrete number of particles. Expectations such as $E(s_t|z_{1:t})$ can be easily computed as Monte Carlo averages:

$$\frac{1}{N}\sum_{i=1}^{N} s_t^i \tag{4.11}$$

In order to instantiate the model we needs: an observation equation $p(z_t|s_t)$ that represent the likelihood function of the observations with respect to the state distribution, a state evolution $p(s_t|s_{t-1})$ that define how the state evolve over time, and an initial state distribution $p(s_0)$.

There are several methods to simulate the sampling of particles, one of these is the *Sequential Importance Resampling* (SIR). This method is composed of three steps:

1. (Propagation) Draw $s_t^i \sim p(s_t|s_{t-1}^i)$ for $i = 1, ..., N$
2. (Resampling) Draw $s_t^i \sim Mult_N(\{w_t^i\}_{i=1}^N)$
3. (Importance normalization) $w_t^i = \frac{p(z_t|s_t^i)}{\sum_{i=1}^N p(z_t|s_t^i)}$

The first step propagates the state of each particle in a new state through the defined state evolution function. The second step resamples the particles using the multivariated distribution defined by the discrete weights at previous time. The third step establishes the new normalized importance values of the weights through the likelihood function.

Since the face detector proposed by Viola and Jones can only produce binary output (face or non-face), it is not possible to determine a likelihood of the measurement system in a probabilistic way in order to fit the particle filter model.

Boccignone et al. [71] propose a methodology to determine a probabilistic model based on the Viola and Jones' face detector. In this work was used the same approach and for this reason we refer to [71] for further information on the implementation of single details. However, to further improve the accuracy of the detector with occlusions and non-linear motion, the evolution of the particles was not modulated with additive Gaussian noise as the article suggests, but a different approach based on optical flow was used.

First the face is detected by the face detector, then the propagation of the particles is made using the speed and direction of the optical flow of the face window predicted at time *t-1*. In this way, unlike Kalman filter, we can give to the particle a non-linear dynamics. Furthermore, we are able to track the motion of the face also if it is partially or completely occluded by hands or other objects, in fact the optical flow allows the evolution of the particles also without the detection of the face, because based only on the assumption of brightness constancy of the pixels over time.

4.3.3 Comparative tests and results

To estimate the quality of the two trackers presented, were made 3 tests. The first test considers a subject moving its head with linear motion, the second test involves a subject moving its head with high speed non-linear motion, the third test investigates a subject moving its head with linear motion but sporadically occluding the face with its hands.

For each test both the tracker processed the three videos and then the number of hits and miss were computed. For hit we intend a window comprehending the face with visible eyes, nose and mouth. The following table shows the results (Tab. 4.1).

	Kalman hits	Particle hits
Test 1	96%	92%
Test 2	93%	97%
Test 3	78%	82%

Tab. 4.1: Results of the comparative tests on the two trackers

In the first test are immediately visible the better performance of Kalman filter with respect to Particle filter. In fact, in addition a higher frame rate, Kalman filter results more stable and has more hits with respect to Particle filter, in details Kalman filter has 96% of hits while Particle filter has 92%.

In the second test it is possible to see the difficulties of Kalman filter on non-linear dynamics. In this case, Kalman filter has 93% of hits due to the loss of most of the frames with abrupt changes of direction, while Particle filter has 97% of hits losing only some of the frame with highest speed.

In the third test Particle filter proves once again superior to Kalman filter. The hits of Kalman filter in this case are 78% with no hits during the occlusion of the face, while the Particle filter has 82% of hits, with only a small portion of frames with occlusions missing. Overall it can be seen that:

1. Kalman filter is optimal for linear dynamics;
2. Both the trackers show a high percentage of hits;
3. The task with the highest number of missing is the tracking of the face partially or completely occluded.

For these reasons the decision was to use Kalman filter at least for this first part of the project, because in controlled situations non-linear dynamics and occlusions are hardly present. However it may be necessary in the future the use of Particle filter and other improvements in order to face real-life problems.

4.4 Facial landmarks localization

The aim of facial landmarks localization is to find the accurate positions of the facial feature points such as the corners of eyes and mouth or the centre of nose (Fig. 4.3) [68].

In the last decades many attempts were made in order to fulfil this task. Early researches extracted facial landmarks based only on geometrical knowledge of the face, for example with the use of contours detection and splines [72]. These model-independent algorithms led to poor results and for this reason the researchers start to focus on model-dependent algorithms. First attempts in this direction were made using rigid face models labelled with facial components [73]. Since these face shape models were not based on statistical learning, their results were once again unsatisfactory.

First successes were due to the Active Shape Model (ASM) [74] and Active Appearance Model (AAM) [75]. In these models, face shape is modelled as a linear

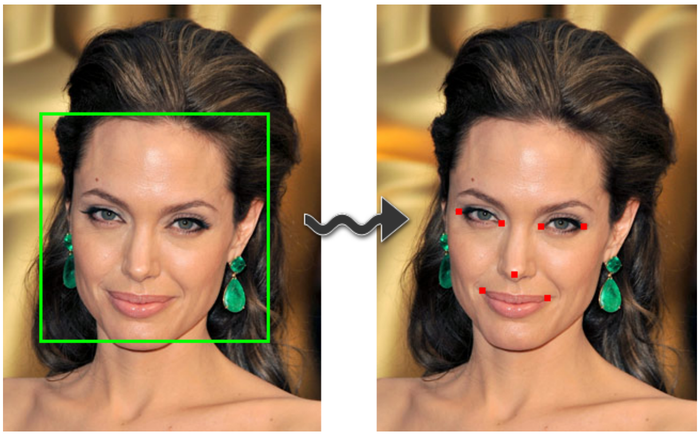


Fig. 4.3: Example of facial landmarks after the detection of a face

combination of principal modes learned by examples of training face shapes in order to learn a deformable shape model through statistical distribution of shape and textures (Fig. 4.4).

With this deformable shape model it is then possible to extract objects with similar shape to those in the training set by fitting the deformable model to images.

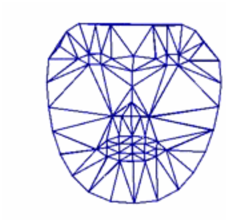


Fig. 4.4: Example of an AAM model for face landmarks localization

For their flexibility and good performance, ASM and AAM became soon the most popular models of facial landmarks detection. However, at now still considerable difficulties are encountered on this task, especially when the face images are in uncontrolled situations. For example, it is difficult for an AAM model to correctly fit both frontal and profile (or semi-profile) faces. The reason could be that the model is still to much rigid to generalize several poses, in fact it concentrates on a global vision of the appearance or shape of the face and not on the subcomponents that compose it.

Deformable Part Models (DPM) take in consideration this issue finding in a first step a match of the whole object, and then, using its part local appearance models, they fine-tune the results minimizing a deformation cost.

The model can be viewed as an undirected graph with a series of vertices relative to the parts of the object and edges between related parts of the object. For example in the case of a face a DPM could be described as a graph in which the vertices are the eyes, the nose and the mouth, and the edges are between eyes and nose, and between nose and mouth (Fig. 4.5). Then, the complexity of this algorithm results related to the structure of the graph, in particular, acyclic graphs allows efficient estimation by Dynamic Programming. With DPM it is possible to fuse the local appearance model and the geometrical constraints into a single model, further improving the quality of the facial landmarks detector.

A work of Uřičář et al. [76] goes in this direction in order to fulfil the detection and localization task of facial landmarks. Uřičář treats the landmarks detection

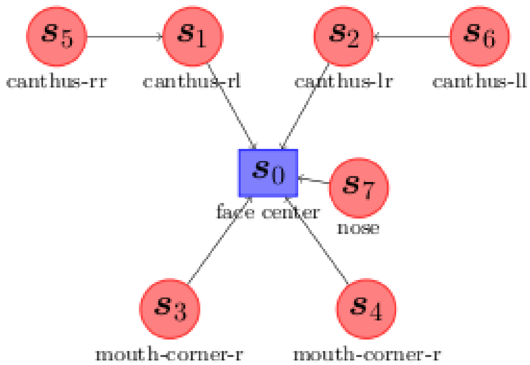


Fig. 4.5: Underlying graph for the landmark configuration

as an instance of the structured output classification problem, using a Structured Output SVM (SO-SVM) [77] as the algorithm for learning the parameters of the landmarks detection from the training set. As appearance model he proposes to use a Local Binary Patterns (LBP) [78] pyramid, that is famous in computer vision for its ability to represent textures and easily allow the detection of similar ones. Finally, as deformation cost function was introduced a quadratic function $g_{ij}(s_i, s_j)$ of displacement vectors $s_j - s_i$ defined as:

$$\begin{cases} \Psi^g_{ij}(s_i, s_j) = (dx, dydx^2, dy^2) \\ (dx, dy) = (x_j, y_j) - (x_i, y_i) \end{cases}$$

In the article were also presented a set of experiments evaluating the performance of the facial landmarks detector. In particular, it is been shown that the performance of this detector based on DPM are better than detector based on AAM. Furthermore, Uřičář released a free version of his detector written in C++.

For these reasons in this project was used the facial landmarks detector proposed by Uřičář et al. [76]. However to further improve the precision of the facial landmarks detector, was added to it a Kalman filter (see Section 4.3.1) in order to decrease the noise on the detection process.

4.5 Facial image normalization

Detecting the faces in the video and tracking their movements are not sufficient in order to extract pictures of faces directly usable to train a regressor; in fact, the face could change the pose or a different illumination could create different colour of the skin, adding noise that may reduce the performance of the regressor. Furthermore, faces have different vertical and horizontal sizes and position of eyes, mouth and nose are not fixed among people.

To solve the first problem of different poses were used the facial landmarks in order to align the face. It was assumed that the faces of interest are only frontal, so every face detected by the profile face detector was discarded and used only as a measure for the face tracker.

Considering a right-handed coordinate system, with the origin at the sensor, Z

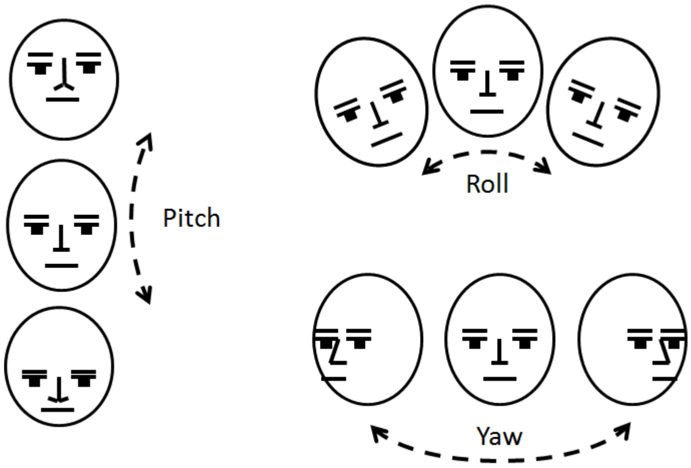


Fig. 4.6: Possible movements of a face around X, Y and Z axis (Pitch, Yaw, Roll)

pointed towards the subject and Y pointed up, a frontal face is able to move around the Z-axis (Roll), around the X-axis (Pitch) and a little around the Y-axis (Yaw) (Fig. 4.6). For each one of these possible face dynamics it was chosen a specific solution.

Normalize the pose of a face rotated around the Z-axis (Roll) it is a quite simple task. It is sufficient to consider the position of the centre of the two eyes in order to estimate the angle ρ between the line passing through the centres of the two eyes and the X-axis of the image (Fig. 4.7).

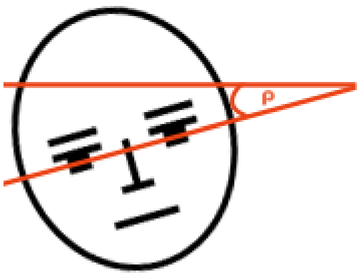


Fig. 4.7: The roll angle ρ

The positions of the centres of the eyes are available thanks to the facial landmarks detector. Using the position of the eyes corners, it is possible to estimate the centres as the mean of the two eyes corners. To estimate the angle ρ it is possible to use a simple equation:

$$\rho = Atan(\frac{e_y^r - e_y^l}{e_x^r - e_x^l}) * \frac{180}{\pi} \tag{4.12}$$

Where (e_x^r , e_y^r) are respectively the x and y coordinates of the centre of the right eye, and (e_x^l , e_y^l) are respectively the x and y coordinates of the centre of the left eye. Given the ρ angle estimated with equation 4.12 it is possible to rotate the face with an affine transformation on the face image.

The normalization of a face rotated around X and Y axis is a more difficult task. In fact it is not only necessary to estimate the rotation angles, but we need also a transformation of a 2D image that simulates a transformation in a 3D space, because with the projection of the 2D image we have lost a dimension and with it important information for the correct reconstruction of the normalized pose.

For this reason we decided to collect information about the current pose of the face and simply discard faces with clues of large rotation angle around X and/or Y

axis.

When a face is frontal, the proportion of the lengths of the two eyes is next to one:

$$\frac{\|e_1^r, e_2^r\|}{\|e_1^l, e_2^l\|} \approx 1 \tag{4.13}$$

A rotation around the Y axis (Yaw) produces a different proportion no more next to one. Observing this behaviour, frames with eyes ratio outside the range $1 - \tau$ and $1 + \tau$ were no more considered. Empirically the τ was set to 0.15.

To detect and then discard faces with pitch was used an eyes and mouth detector, based again on Haar feature and cascade of strong classifiers trained this time with pictures of eyes and mouths. Since the classifier was trained with frontal images with a less degree of rotation, faces with a large pitch angle do not give sufficient clues to the detector in order to detect the eyes and/or the mouth of the subject. Pictures that do not pass this test were discarded.

This last step it is important not only for discarding faces with large pitch angle, but also to discard pictures not containing correct faces with clear eyes and mouth, that are necessary in order to estimate the affective state of the subject.

When a face picture pass all these tests, it is necessary to normalize the size of the window and the relative position of the facial landmarks.

The face window was fixed to 90x110 pixels and the image was scaled in order to normalize the eyes baseline of the subject was to 70 pixels of length and the vertical distance between eyes and nose to 31 pixels (Fig. 4.8). Then the window was cropped around the face. In this way we are able to obtain pictures of faces with normalized proportions.

The next step is the illumination normalization. To fulfil this task was used a particularly robust algorithm proposed by Tan and Triggs as a pre-processing step for a face recognition application [79].

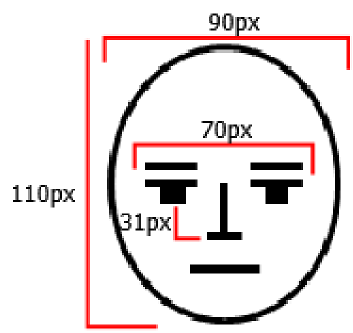


Fig. 4.8: Normalized sizes of the face

The first step is to apply a *gamma correction* on the gray-level image I. The gamma correction process provide a non-linear transformation that replaces I with I^γ with $\gamma > 0$. The effect is the enhancement of the local dynamic range of the image in dark or shadowed regions and the compression of it in brighter regions. The author suggests a gamma with value $\gamma = 0.2$.

The second step uses a *difference of Gaussian* (DoG) as a filtering to remove shading effects; in fact the gamma correction is not sufficient in order to remove the shadings on the face texture. Using a DoG it is possible to implement a band-pass filter that removes low frequency information, such as the shadings, and high frequency details such as the noise on the image. The problem is to determine how much have to be wide the inner band. The author suggests to use $\sigma_0 = 1.0$ and $\sigma_1 = 2.0$ as values of the sigma relative to the two Gaussians.

The final step globally rescales the image intensities to standardize a robust measure of overall contrast. To accomplish this task it is necessary to take in account that the image could contain small portions of extreme values due to highlights, garbage at the image borders and small dark regions such as nostrils, so it is crucial to use an estimator robust to this kind of noise. It is eventually possible to use a mask for useless portion of the image. The author suggests a two stage process:

$$I(x,y) \leftarrow \frac{I(x,y)}{(avg(|I(x,y)|^{\alpha}))^{\frac{1}{\alpha}}} \tag{4.14}$$

$$I(x,y) \leftarrow \frac{I(x,y)}{(avg(min(\tau,|I(x,y)|)^{\alpha}))^{\frac{1}{\alpha}}} \tag{4.15}$$

Where α is a strongly compressive exponent reducing the influences of large values and τ is a threshold with the function of truncating large values after the first step of the normalization. The values suggested are $\alpha = 0.1$ and $\tau = 10$.

To further remove extreme values still contained after the normalization process, was used a hyperbolic tangent to compress values to the range $(-\tau, \tau)$:

$$I(x,y) \leftarrow \tau tanh(\frac{I(x,y)}{\tau}) \tag{4.16}$$

Then a scaling of the values of the image pixels is done in order to have the values in the range $(0, 1)$.

Chapter 5

Preliminary tests and results

Science is nothing but perception

Plato

5.1 Introduction to tests

The aim of this chapter is to investigate with some preliminary tests the validity of the choices made in the previous chapters. Our approach is quite innovative and overturns the vision of the present researches on emotions recognition. In fact whereas the majority of present researches focus on the idea that the observation with a specific affective power generates an affective latent variable, our vision investigates the opposite behaviour, namely it is the affective latent variable that possesses a specific affective power and consequently generates a congruous observation.

For this reason, the current datasets for emotions recognition present labelled observations, where these labels could be seen as the representation of observations' affective power. Although with these datasets it is possible to regress a latent space with several supervised machine learning techniques, the topology of the resulted space is biased by the representation used to describe the emotional power of the observations in the dataset. Nevertheless in previous chapters we told several times that currently there is not a shared theory among the psychologists about the representation of emotions, so this bias could be simply wrong.

Our idea allows to solve this issue, since the method used for the regression of the latent space is not supervised, but based on an *unsupervised* probabilistic dimensionality reduction. However, the downside of this approach is that objective evaluations of the performance are harder to make, for the reason that the topology of the latent space is described by a set of not always clear axes, whereas the observations used for the training process are usually described by a set of well defined labels, so that each attempt to compare them becomes a difficult task.

There are at least two ways to make objective evaluations of the performance with our approach: the first involves the use of several subjects evaluating the similarity between the observations in the dataset and the observations generated by the latent space with a likert scale or similar approaches; the second requires a training set and

a test set where each observation is labelled with specific characteristics of the facial expression in order to allow one-to-one relationships between the two sets, which could be used to estimate the distance in the latent space between the position of the observation in the training set with that in the test set.

The first method presents as advantage the possibility to leave the evaluation process to human subjects representing the future users of the application and having first important information on the ability of the model to recognize emotions as humans do. The disadvantage is that, in order to produce statistically significant evaluations, many subjects are needed differing in ages, sex, cultures, education, and so on, which is not always simple to arrange, at least not in a short period of time like that available to develop this preliminary work.

The second methodology has as its major pro the ability to produce numerical, sound and objective results; however it is not easy to have datasets with a one-to-one relationship between the two sets of observations. For example this is not possible when the labels comprehend several shades of the concept that they describe (e.g. the classes of the basic emotions). Furthermore, when the dataset has this characteristic, it is still possible to have misalignments between the two facial expressions under evaluation, due to the different evolution of these over different times.

This thesis focuses on a *preliminary* observation of the classification behaviour of the proposed model based on GP-LVM. For this reason the following tests have the sole purpose to give us high level information to determine whether or not the proposed approach could be spent in future works, providing some useful guidelines for its potential improvements. As consequently the results we achieve are not sufficiently relevant for a scientific and sound evaluation, we will devote future works to improve their quality.

For the following tests and their evaluations it was used the MATLAB® code[1] created by Lawrence with several useful functions for the use of GP-LVM. This algorithm needs to specify the number of active points, namely the number of points selected by the IVM algorithm, the number of maximum cycles of optimization, and obviously the desired dimensionality of the latent space.

5.1.1 Datasets and issues concerning data collection

Major difficulty for affective computing applications is to collect good quality data in order to use it for the training process and then to evaluate the performance. This difficulty is due to several causes, most of which were already told in previous sections. However for greater clarity they will be listed above as a brief summary:

The *concept of emotion* under investigation – There are several concepts compatible with the term emotion. For example we can intend the current internal affective state of the user, a long term dynamics of affective states or simply the external feedbacks given by the subject in a precise moment. In the first two cases it is difficult to produce good data quality, for the reason that the only way to try to measure the internal affective state of a user is to place on it annoying and invasive

[1] http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/fgplvm/

sensors, which augment the risk of biased data, and consequently to infer the affective state [80]. Furthermore current sensors are not so precise, therefore the data are very noisy. In the latter case the data collection is simpler, but not sufficient easy. In fact we can use non-invasive sensors like a camera and a microphone; however it is still possible to collect noisy data due to the insufficient quality of the sensors or experimental conditions, especially when data are collected in uncontrolled situations.

Acted VS. non-acted emotions – It is possible to choose between acted and non-acted emotions. In the first case a subject is asked to produce a particular facial expression and/or speech in order to stress a particular emotion. In the second case the external signals given by the user are more natural, because they are induced with particular techniques or are filmed in secret during a particular social interaction. The former are simpler to collect but also poor of information, for the reason that often the expressions are unnatural and/or not very prominent. Furthermore acted data usually do not include the full range of possible emotions available in a common social interaction. The latter are more informative being data collected from real life, but they suffer of a severe noise, due for example of different face poses, non-uniform light conditions and lips movements during the speech that could produce fake emotional data.

Labelling process – The labelling process is not a simple task; determine a label for a particular affective state is usually subjective. More objective methods of labelling imply rigid classes and pattern schemas, causing the creation of datasets often unsuitable for most real life applications. A good labelling approach should provide an objective labelling procedure maintaining in the same time a wide range of possible schemas.

Taking into account these difficulties, the data were initially collected from a set of videos available on Youtube[2]. This collection includes six subjects, three males and three females, filmed during a real interview or during an acted[3] monologue and covering a wide range of affective states and facial expressions. With this choice, it was possible to evaluate performance directly on real life data; however labels were not present, consequently this issue created difficulties during the evaluation process.

Considering the problems of the previous dataset, it was contemplated the use of a laboratory dataset. For several reasons our choice was oriented on the MMI-Facial Expression Database collected by Valstar and Pantic [81]. The power of this dataset is that almost each frame of the videos in it is labelled with the Action Units present on the face of the subject, and, even if the facial expression is acted, it is possible to cover a wide range of facial expressions (although only single or simple combinations of them were considered for each video).

[2]http://www.youtube.com

[3]Acted but not constrained, so the facial expressions appear to be highly naturalistic and spontaneous

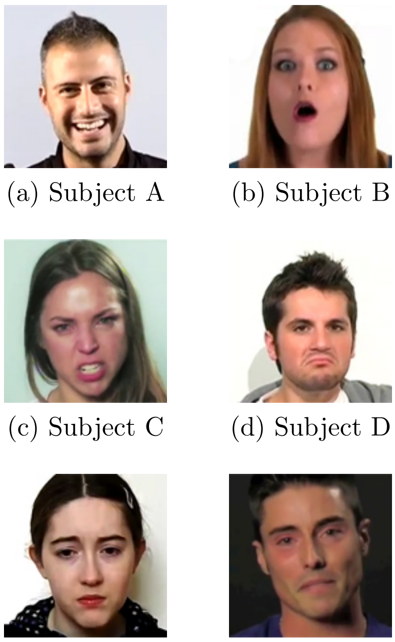
	anger	disgust	fear	happiness	neutral	sadness	surprise
Subject A	✗	✗	✗	✓	✓	✗	✗
Subject B	✗	✓	✓	✓	✓	✗	✓
Subject C	✓	✗	✗	✗	✓	✓	✗
Subject D	✗	✗	✗	✓	✓	✗	✓
Subject E	✓	✗	✗	✗	✓	✓	✗
Subject E	✗	✗	✗	✓	✓	✓	✗

Tab. 5.1: Details of the web-video dataset

5.2 Dataset of heterogeneous subjects from Web

The Web contains a lot of information: most of them are unfortunately unstructured or semi-structured; nevertheless the quantity of data available in it is often superior to that possibly collected in a laboratory.

Our idea involves the collection of several videos from Youtube and consequently the extraction of the included facial expressions to train the model. This decision born from the conviction that the stronger deficit of actual facial expressions databases is the unreality of facial expressions in them, the absence of important social signals, and often the aseptic character of the included facial expressions. Accordingly, collecting a set of videos of real interviews or highly spontaneous monologues it seemed to us a good choice to stress the emotional character of the faces.



(a) Subject A (b) Subject B (c) Subject C (d) Subject D (e) Subject E (f) Subject F

Fig. 5.1: Subjects of the web-video datasets

The dataset includes six subjects: three males and three females (Fig. 5.1). Each video includes several facial expressions; therefore to cover all the range of emotions we tried to collect faces exhibiting all the six basic emotions and their shadings. Unfortunately the footages containing emotions like disgust and fear were difficult to find and for this reason they are present only in a small portion of frames. In Tab. 5.1 we summarize the emotions expressed by the subjects; obviously the classification of the frames into these six basic classes of emotions is reductive, but it was necessary as a guideline in order to create a rather balanced dataset.

The major difficulty for this kind of dataset is that the environment conditions were not controlled, so light conditions can change during the shot and also the pose of the face. Furthermore some facial occlusions are sporadically present.

Another problem is that the footages are long, and this produces several frames for each subject, causing on the one hand a dimensionality problem (the memory

Fig. 5.3: Examples of unnatural faces generated from latent space's points

necessary to process the whole data) coupled with a lot of redundancy, and on the other hand a problem for the regression with GP-LVM, since this model is known to produce good regressions with training set of small dimension [82]. To solve this problem, for each subject, the 150 most informative frames were sampled with the use of the IVM algorithm presented in Section 3.5. After this sampling process our dataset included 900 frames in total.

5.2.1 Pilot test

The aim of this first test was to verify the properties of the latent space regressed with GP-LVM after few cycles and with the use of a small amount of active points in order to get familiar with this dimensionality reduction tool.

The number of maximum cycles and the number of active points were set to 100 and the dimensionality of the latent space to 2. The latent space resulted is shown in Fig. 5.2, where the identity of each subject in the training set is drawn in a different colour and/or shape.

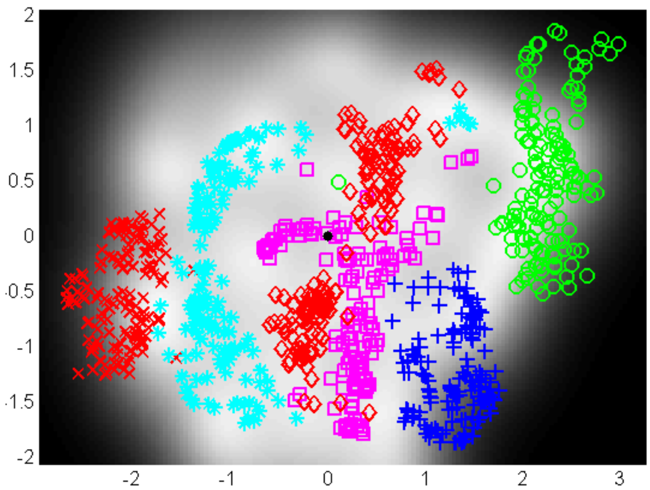


Fig. 5.2: The latent space generated after 100 cycle and using 100 active points

The resulted latent space shows a rather well separated identities of the subjects, even if it was not our aim. In fact our need is to separate different facial expressions, and, considering that a similar facial expression is usually produced by several subjects in the dataset, this imply that the latent space should present the identities more mixed up in order to fulfil this aim. Furthermore, the points of the latent space generates noisy observations of the data space, resulting in unnatural faces and causing non-smoothing dynamics (Fig. 5.3).

To solve the first problem a possible solution is to augment the number of active points used during the optimization process in order both to have more sparse facial expressions and to augment the chances of generating clusters of them; whereas to solve the second problem it is surely convenient to highly augment the number of

Fig. 5.5: Examples of classification of new observations. In the first row the new observed data are reported, whereas the second row contains the data generated by the model according to the likelihood of the new observation w.r.t. the latent points.

cycles.

5.2.2 Varying the number of cycles and active points

In this test the number of cycles was augmented to 1000 and the number of active points to 300. The resulted latent space is shown in Fig. 5.4.

The identities are further grouped into well separated clusters, which makes difficult to regress the similar facial expressions among the subjects of the dataset. This means that augmenting active points is not the right answer to this problem.

On the contrary, the noise on the observations generated by the latent points was drastically reduced, causing smooth dynamics among close points of the space and more realistic faces.

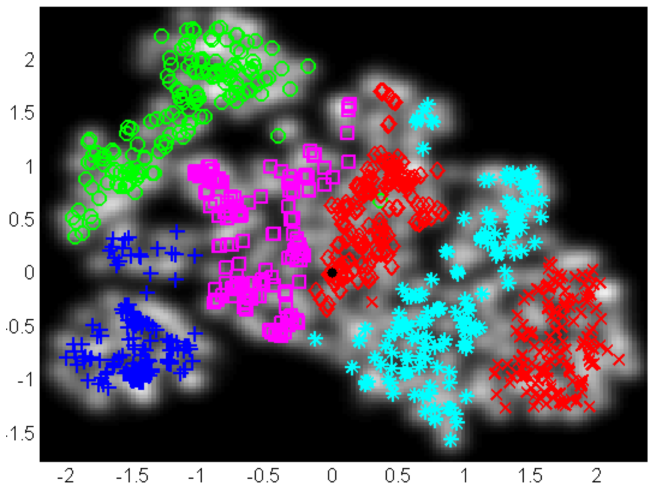


Fig. 5.4: The latent space generated after 1000 cycle and using 300 active points

To verify the ability of the model on classifying new observations, we sampled 10 facial expressions from another video available on the Web containing facial expressions similar to those used for the training process, but with different subjects. Each new observation was classified by the model according to the likelihood of the new data w.r.t. the latent variables. The results are shown in Fig. 5.5.

It is possible to qualitative see that the classification process fails and that it suffers from a bias on the pose of the face, especially the face shape. These results confirm that the identity, and consequently the dimension and shape of the face, drastically reduces the ability of the model to classify new observations.

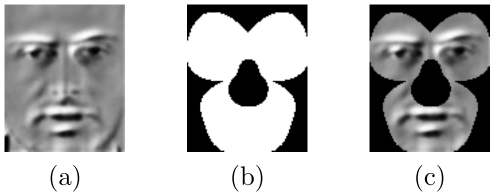
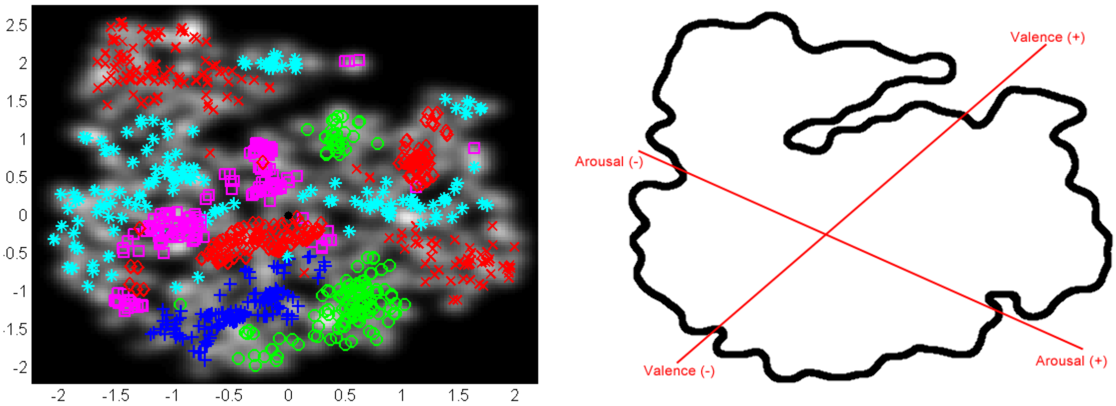


Fig. 5.6: Example of application of the mask on a subject. In (a) the frame before applying the mask, in (b) the mask and in (c) the frame after the application of the mask.

5.2.3 The identity problem

To solve the problem occurred in the previous test (Section 5.2.2) the only way is to hide the useless parts of the faces that could produce a bias on the identity of the subjects. A possible approach is to use as features for the classification only the positions of facial landmarks and/or splines describing the shape of mouth and eyes. However, in our opinion, this approach is not sufficient to cover the complex structure of a non-trivial facial expression, since information such as wrinkles would be lost.

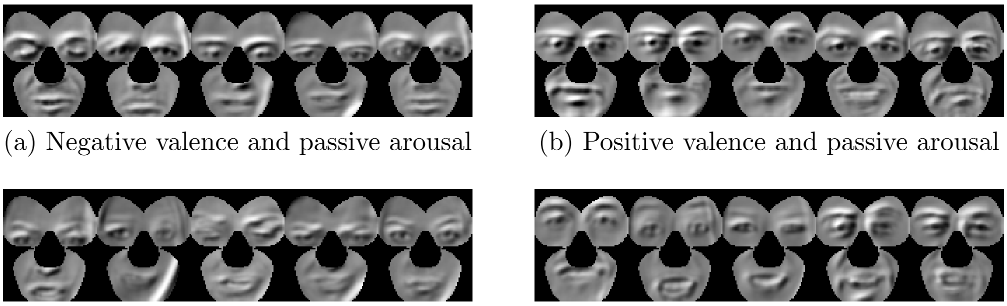
A second possible scenario is to cover with a mask the pixels of the face useless for emotion recognition, trying to hide as more as possible the identity of the subject without occluding important cues of its facial expression. To fulfil this task the mask in Fig. 5.6 was proposed and a new model was trained with only the pixels not covered by the mask. Also in this case the cycles was set to 1000 and the active points to 300, regressing the latent space in Fig. 5.7 (a).



(a) (b)

Fig. 5.7: The latent space generated after 1000 cycle with masked frames (a) and the relative topology accordingly to that of core affect proposed by Russell (b)

This new latent space has the identities more mixed up. By investigating the dynamics among close points in the latent space smoother dynamics in the facial expressions are observed, even if the identity changes during the movement on the selected path.



(a) Negative valence and passive arousal

(b) Positive valence and passive arousal

(c) Negative valence and active arousal

(d) Positive valence and active arousal

Fig. 5.8: Samples from the four quadrants of the generated core affect

Furthermore the latent space shows an interesting topology. It seems that there is a correlation between the core affect space theorized by Russell (cfr., Section 2.2.6) and our space generated by GP-LVM, since the four quadrants simplified in Fig. 5.7 (b) are visible like those in Fig. 3.1.
Fig. 5.8 shows 5 observations for each area of the regressed core affect space. Anyway the subdivision in four areas is not always well distinct, as proved by the almost non-linear boundaries of the arousal. Conversely the valence is more precisely divided.

To test the classification performance on this new space we used the same observations of the previous test (cfr., Section 5.2.2). The results are shown in Fig. 5.9.

Although still not satisfactory, the current results are more accurate than the previous test's. The problem was largely due to the quality of the dataset. First of all the number of facial expressions present in the dataset is high, but not sufficient to cover the whole range of emotions. This in turn is due to the difficult task of collecting videos of single subjects exhibiting a series of facial expressions without large changes of pose and with a good quality of the footage. The second problem is related to the repetition of the facial expressions in the dataset: the different identities cause a separation over the space of these similar facial expressions even if they should be placed in close areas.

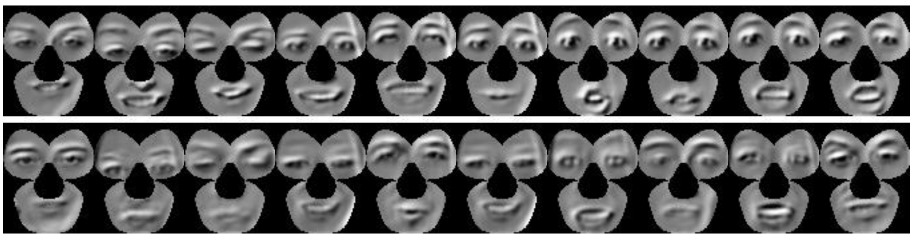


Fig. 5.9: Examples of classification of new observations. The first row reports the new observed data, whereas the second row contains those generated by the model according to the likelihood of the new observation w.r.t. the latent points.

(a) Subject 0 (b) Subject 1

Fig. 5.10: The subject of MMIDB used for the tests

5.3 MMI Facial Expression Database

The MMI-Facial Expression Database [81] contains more than 1500 samples of both static images and image sequences of frontal and profile faces exhibiting various facial expressions, single Action Unit activation, and multiple Action Units activation. An Action Unit (AU) describes the activation of a single facial muscle as defined in the Facial Action Coding System (FACS) [54]. FACS is a system designed with the aim of giving an objective description of a facial expression changes in term of observable facial muscle actions (Fig. 5.11). This system provides rules for the visual detection of 44 different AUs and the relative temporal segments (onset, apex, offset).

The subjects included in the database are 19 and they have different gender, race, age and facial characteristics such as beard, glasses and moustache. Each subject produced a series of footage each containing either a single AU or a combination of a minimal number of AUs (when for instance a single AU cannot be displayed alone).

Fig. 5.11: Examples of Action Units

The database includes both frontal and profile faces; the subjects were asked to display the required expressions while minimizing out-of-plane head motions.

Most of the frames in the database was described in terms of displayed AUs; moreover the access to the database is web-based with the possibility of filtering the data with a specific query[4].

For our purpose two subjects were used: the first (Subject 0) for the training process and the second (Subject 1) for the test process (Fig. 5.10). To reduce redundancy and dimensionality of the frames of Subject 0, we subsampled each footage selecting only the frames with the AU on apex. Consequently, considering that these frames represents a small fraction of the total, the training set size decreased drastically to approximately 900 frames (Fig. 5.12).

[4]http://www.mmifacedb.com/

Fig. 5.12: Examples of facial expressions contained in the dataset selected for the training process

5.3.1 Pilot test

As for the previous dataset, the aim of this first test was to verify the regression ability of GP-LVM with the dataset collected from the MMI database. Since the actual and the previous datasets have more or less the same amount of training data, the number of cycles was set to 1000, the number of active points to 300 and the dimensionality of the latent space to 2. The resulting space is shown in Fig. 5.13 where the different combinations of AUs are drawn with different colours and/or symbols.

From Fig. 5.13 it is evident that some AUs combinations are well separated in the space, whereas others are mixed up in a single area. Due to the large number of classes (52), the specific AUs not well separated by GP-LVM cannot be visually distinguished. However a more deep investigation shows how the classes well grouped into clusters include more prominent facial expressions (Fig. 5.14). Unfortunately, most of the AUs in the dataset are similar each other as the changes of the facial expressions w.r.t. a neutral one are very small. Consequently it is difficult for GP-LVM to assign them to well distinct classes in the latent space.

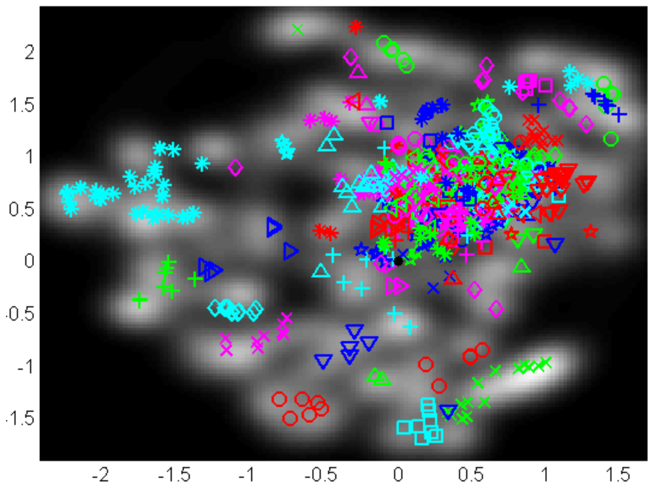


Fig. 5.13: The latent space generated after 1000 cycles

Unlike the latent space generated from the web dataset faces, the topology of this latent space gives no information on arousal, valence or other affective variables.

Fig. 5.14: Examples of faces belonging to classes well separated in the latent space generated by GP-LVM

Concerning the classification ability, unlike the previous tests, we used here most of the AUs performed by Subject 1 and also some of the most characterizing frames of the subjects appearing in the web dataset. The results are shown in Fig. 5.16 and 5.17 respectively.

A qualitative analysis of the results shows a greater accuracy on web dataset faces: the more prominent character of these facial expressions allows them to fall into well-divided classes in the regressed latent space. On the contrary, the MMI dataset do not present facial expressions strong enough, displaying movements of the sole facial muscle under exam and showing consequently a little – or even null – emotional power.

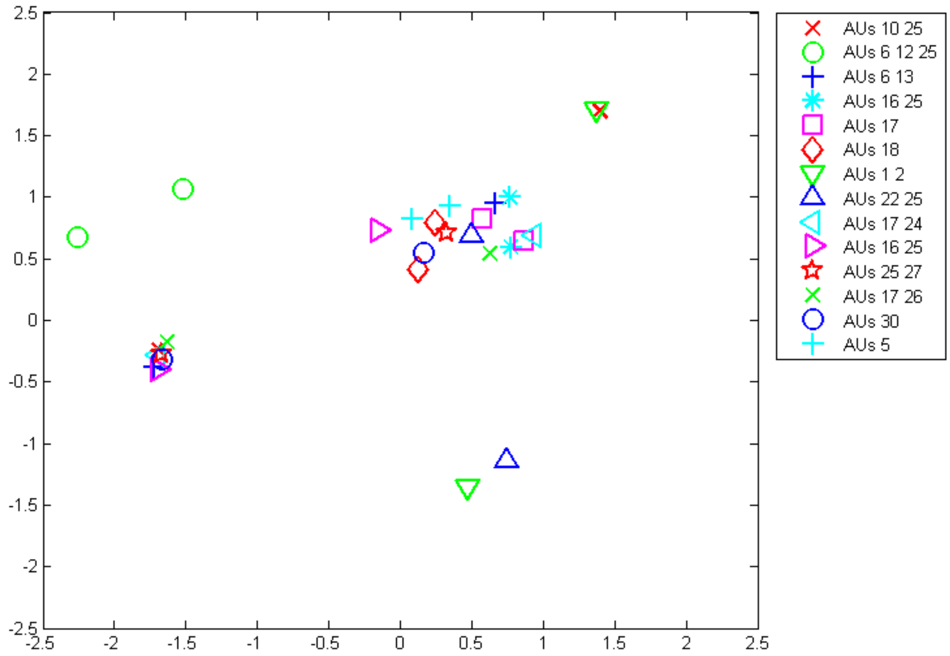


Fig. 5.15: Positions of true and test faces in the 2D latent space

For some of the faces of Subject 1 it was possible to compute the distance relative to the position of the face of Subject 0 activating the same set of AUs. Unfortunately, the presence of unlabelled faces in the dataset and discrepancies between the AUs activated by Subject 0 and those activated by Subject 1 prevents us from accomplishing this task for all the faces.

The results of the classification in the latent space are shown in Fig. 5.15 where each combination of AUs is represented with a different colour and/or shape.

In the picture some sets of AUs were classified close to the relative true position

Pred. \ Real	AUs 10 25	AUs 6 12 25	AUs 6 13	AUs 16 25	AUs 17	AUs 18	AUs 1 2	AUs 22 25	AUs 17 24	AUs 16 25	AUs 25 27	AUs 17 26	AUs 30	AUs 5
AUs 10 25	**0.79**	0.83	0.81	0.28	0.26	0.32	<u>0</u>	0.63	0.8	0.81	0.79	0.77	0.8	0.34
AUs 6 12 25	0.29	<u>**0.18**</u>	0.32	0.51	0.53	0.39	0.65	0.69	0.3	0.32	0.29	0.27	0.3	0.35
AUs 6 13	0.57	0.64	**0.59**	<u>0.08</u>	<u>0.08</u>	0.1	0.23	0.46	0.58	0.59	0.57	0.56	0.58	0.13
AUs 16 25	0.6	0.66	0.62	**0.09**	<u>0.08</u>	0.12	0.2	0.47	0.6	0.62	0.6	0.58	0.6	0.15
AUs 17	0.54	0.62	0.56	<u>0.07</u>	<u>**0.07**</u>	<u>0</u>.07	0.26	0.43	0.55	0.56	0.54	0.53	0.55	0.11
AUs 18	0.42	0.52	0.44	0.15	0.17	<u>**0.09**</u>	0.39	0.36	0.42	0.43	0.42	0.4	0.42	<u>0</u>.09
AUs 1 2	0.53	0.74	0.52	0.43	0.44	0.47	**0.69**	<u>0.07</u>	0.53	0.51	0.52	0.52	0.52	0.48
AUs 22 25	0.52	0.6	0.54	<u>0.06</u>	0.08	<u>0.06</u>	0.29	**0.4**	0.52	0.53	0.52	0.5	0.52	0.1
AUs 17 24	0.61	0.7	0.62	0.04	<u>0.02</u>	0.15	0.24	0.4	**0.61**	0.62	0.6	0.59	0.61	0.19
AUs 16 25	0.39	0.46	0.42	0.2	0.22	0.09	0.4	0.45	0.4	**0.41**	0.39	0.38	0.4	<u>0.06</u>
AUs 25 27	0.49	0.56	0.5	0.1	0.12	<u>0.03</u>	0.32	0.41	0.49	0.5	**0.48**	0.47	0.49	0.06
AUs 17 26	0.53	0.63	0.55	<u>0.03</u>	0.06	0.1	0.3	0.37	0.54	0.54	0.53	**0.52**	0.53	0.13
AUs 30	0.44	0.53	0.46	0.13	0.15	<u>0.06</u>	0.37	0.39	0.44	0.45	0.44	0.42	**0.44**	<u>0.06</u>
AUs 5	0.51	0.57	0.53	0.12	0.13	<u>0.04</u>	0.28	0.46	0.52	0.53	0.51	0.49	0.52	**0.06**

Tab. 5.2: Normalized distances among the training and test points in 2D latent space

(AUs 6 12 25, AUs 16 25, AUs 17, AUs 18, AUs 5), whereas others were collocated far away in the whole space (AUs 10 25, AUs 6 13, AUs 1 2, AUs 22 25, AUs 17 24, AUs 16 25, AUs 25 27, AUs 17 26, AUs 30).

In order to obtain a measure of the classification accuracy, we used the Euclidean distance between the predicted position and the true one. The distances were normalized to the maximum distance among all the points of the training set in the latent space. The results are shown in Tab. 5.2 where the minimum distances are underlined and the distances between predicted and true position of the same AUs are in bold. The numerical results confirm the qualitative analysis previously made.

To gain further information on the classification accuracy of the model, we propose in Tab. 5.3 the following set of descriptive statistics, namely: the mean, median, standard deviation, quantile 0.1 and quantile 0.9.

Mean	Median	Standard deviation	Quantile 0.1	Quantile 0.9
0.3888	0.4273	0.2489	0.0734	0.7038

Tab. 5.3: Results of the classification test in 2D latent space

From the above analysis, it is quite evident that these results are not brilliant, anyway the tests are preliminary and there is room for future improvements.

5.3.2 Augmenting the dimensionality of latent space

The problem presented in the previous section could probably be solved by enhancing the classification ability of GP-LVM in order to have the classes more divided from each other in the latent space. Unfortunately, the increasing of the number of cycles does not enhance the classification ability of the model, as in this way only a reduction of the noise on the observations generated from the latent space is guaranteed.

A possible solution could be to augment the dimensionality of the latent space in order to allow the model to gain in classification accuracy thanks to the information subtended in the new dimensions.

We first tried to regress a three-dimensional space iterating the process for 1000 cycles and using 300 active points; however the observations generated by the latent points were affected by noise.

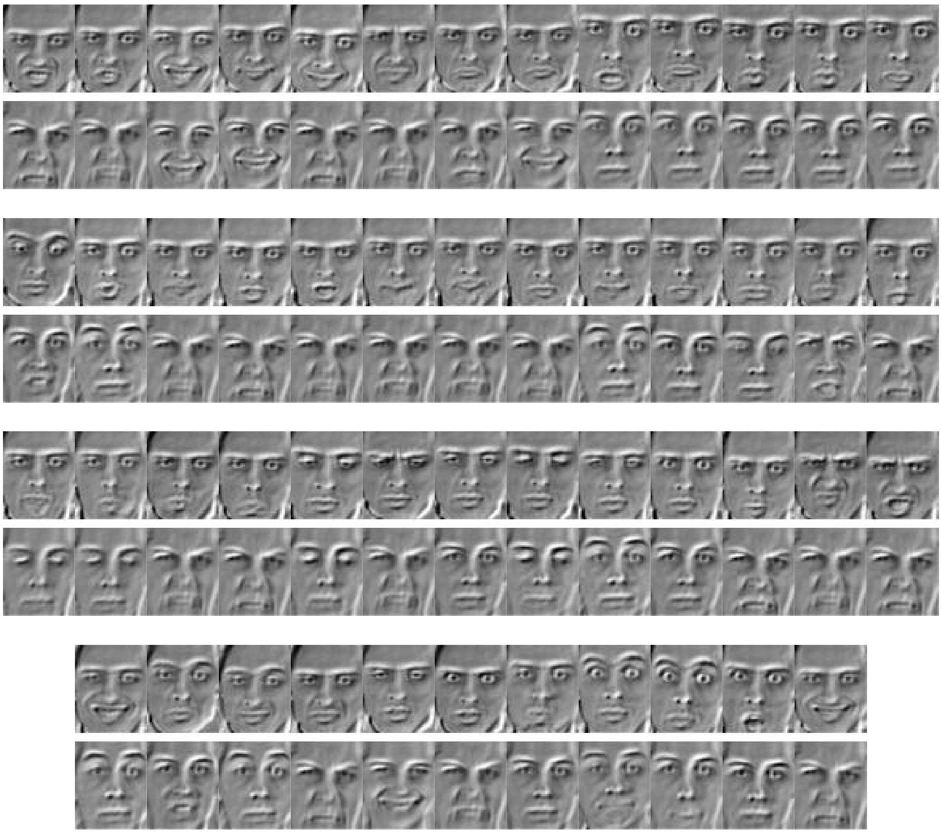


Fig. 5.16: Classification results of Subject1 facial expressions

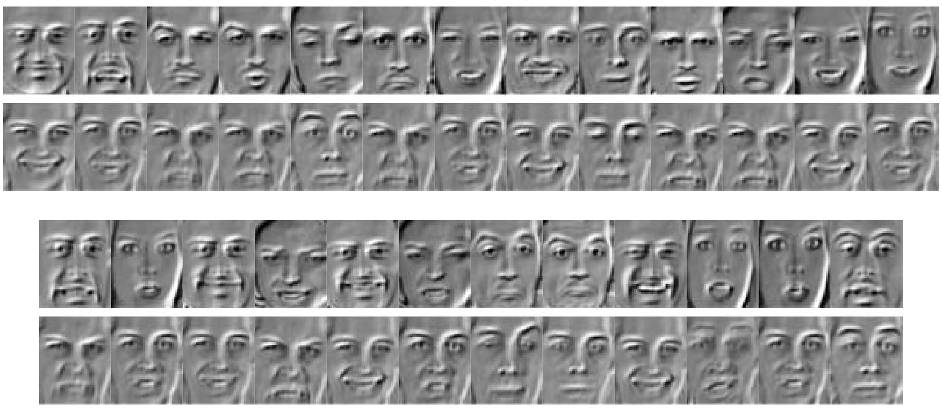
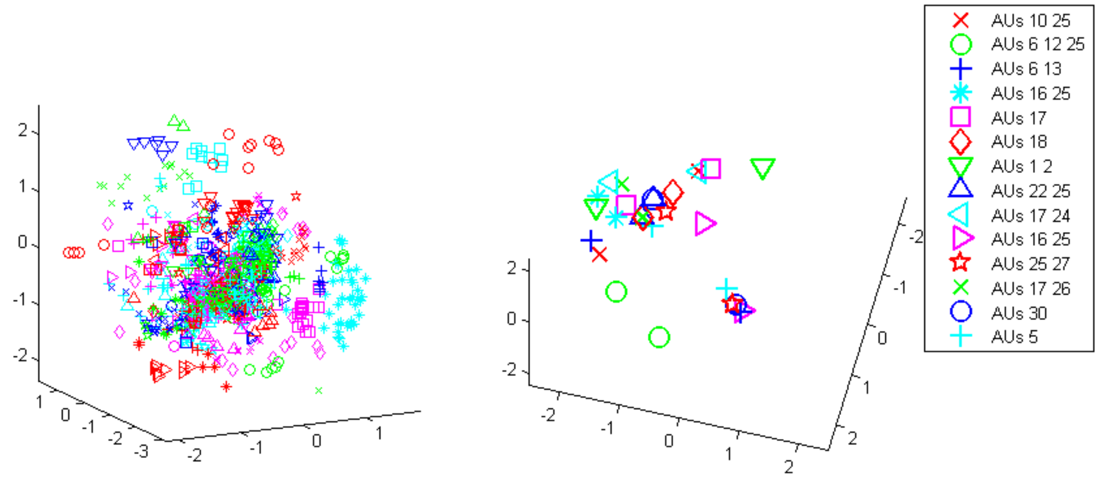


Fig. 5.17: Classification results of faces from the web dataset



(a) (b)

Fig. 5.18: In (a) is presented the latent space generated after 3000 cycle in a 3D space, whereas in (b) are the positions of true and test faces in the 3D latent space.

To check whether the chosen maximum number of cycles was insufficient to achieve satisfactory results, we augmented it to 3000 maintaining the same number of active points. The resulting space after 3000 cycles is shown in Fig. 5.18 (a). Here again we observe the same problem affecting the 2D latent space, namely the intrinsic difficulty to separate the same classes with a not so prominent emotional power.

For the classification test we used the same test faces of the previous test. The results shown in Fig. 5.19 and Fig. 5.20 are not encouraging, being sometimes even worst than in the 2D space. To deeply investigate the accuracy of the model, we show the position of predicted and true observation in Fig. 5.18 (b) and the relative normalized distances in Tab. 5.4.

At a first glance, results in Tab. 5.4 are better than those obtained from the 2D latent space; however the data are less informative than those in the previous tests. It is possible to transform the distances of each row in probabilities of belonging to

Real / Pred.	AUs 10 25	AUs 6 12 25	AUs 6 13	AUs 16 25	AUs 17	AUs 18	AUs 1 2	AUs 22 25	AUs 17 24	AUs 16 25	AUs 25 27	AUs 17 26	AUs 30	AUs 5
AUs 10 25	**0.4**	0.37	0.48	0.19	0.44	0.24	0.15	0.24	0.4	0.47	0.44	0.24	0.45	0.4
AUs 6 12 25	0.49	**0.2**	0.39	0.34	0.53	0.25	0.32	0.25	0.5	0.39	0.37	0.25	0.37	0.36
AUs 6 13	0.44	0.37	**0.51**	0.21	0.48	0.17	0.2	0.17	0.44	0.51	0.48	0.17	0.49	0.46
AUs 16 25	0.32	0.41	0.48	**0.12**	0.36	0.1	0.12	0.1	0.32	0.48	0.45	0.1	0.45	0.41
AUs 17	0.33	0.43	0.5	0.2	**0.37**	0.08	0.21	0.07	0.33	0.49	0.47	0.07	0.47	0.44
AUs 18	0.29	0.47	0.46	0.32	0.33	**0.15**	0.34	0.15	0.3	0.45	0.44	0.15	0.44	0.41
AUs 1 2	0.23	0.71	0.53	0.53	0.18	0.5	**0.54**	0.5	0.22	0.52	0.5	0.5	0.5	0.44
AUs 22 25	0.24	0.44	0.45	0.22	0.28	0.07	0.23	**0.07**	0.24	0.45	0.42	0.07	0.42	0.38
AUs 17 24	0.27	0.55	0.58	0.07	0.31	0.23	0.09	0.23	**0.27**	0.58	0.54	0.23	0.55	0.49
AUs 16 25	0.37	0.4	0.36	0.43	0.39	0.23	0.44	0.23	0.37	**0.36**	0.35	0.23	0.35	0.35
AUs 25 27	0.27	0.4	0.4	0.27	0.3	0.08	0.28	0.08	0.27	0.4	0.37	**0.08**	0.38	0.34
AUs 17 26	0.31	0.5	0.55	0.17	0.35	0.13	0.19	0.12	0.31	0.55	0.52	**0.12**	0.52	0.48
AUs 30	0.32	0.44	0.47	0.28	0.36	0.11	0.3	0.11	0.32	0.47	0.45	0.11	**0.45**	0.43
AUs 5	0.3	0.35	0.39	0.25	0.34	0.04	0.25	0.04	0.3	0.38	0.36	0.04	0.36	**0.33**

Tab. 5.4: Normalized distances among the training and test points in 3D latent space

Mean	Median	Standard deviation	Quantile 0.1	Quantile 0.9
0.3050	0.3441	0.1514	0.1181	0.5126

Tab. 5.5: Results of the classification test in 3D latent space

a particular facial configuration using the following equation:

$$\begin{array}{rcl} \hat{p}(i,j) & = & 1 - d(i,j) \\ p(i,j) & = & \frac{1}{\sum_{k=1}^{J} \hat{p}(i,k)} \hat{p}(i,j) \end{array} \tag{5.1}$$

Where $d(i,j)$ is the distance in row i and column j. The result is a confusion matrix, which is possible to use for computing the conditioned entropy $H(X|C)$ as:

$$\begin{array}{rcl} X_c & = & \{p(c,j_1), ..., p(c,j_n)\} \\ H(X_c|C) & = & \sum_{k=1}^{J} X_{ck} \log_2(\frac{1}{X_{ck}}) \end{array} \tag{5.2}$$

Using Eq. 5.2 on Tab. 5.2 and Tab. 5.4 we obtain as entropies ≈ 0.2 and ≈ 3.7 rispectively, which confirms that the results of the previous test are much more informative than those of the current test.

The points in this new latent space are more concentrated in specific areas; consequently the points belonging to a specific area have smaller distances among each other causing the generation of a facial expression not sufficiently similar to the AUs configuration under exam. This in turn is due to the fact that the generated observation is a mean; therefore if the points of a specific class are not well divided from other classes, the generated observation will be not congruous with what expected.

In Tab. 5.5 we report the mean, median, standard deviation, quantile 0.1 and quantile 0.9 of the previous results.

What we learnt is that the identity problem is not the unique issue in collecting emotional data for training a GP-LVM. In fact it is necessary also that the faces exhibit prominent facial expressions with some informative emotional power in order to generate a space where the classes are well-separated. Without this characteristic the classification process remains challenging with GP-LVM, at least using as features the whole pixels of the face.

Probably, the intrinsic dimension of the latent space is higher than those tested here. This is due to the fact that changes of poses, lightness (when the light conditions normalization fails) and other distortions on the training data increase this intrinsic dimension of the latent space, which consequently is not limited to changes of the facial configuration. Since sometimes the noise on the observations is stronger than the changes of the facial configuration, the model probably considers the AUs changes as less important dimensions, discarding them during the process of dimensionality reduction. Conversely, the noise due to the face pose and other similar distortions is wrongly considered a more important feature, so that it affect the resulting latent space.

Differently, the results obtained with the web dataset are encouraging. As they are real facial expressions caught during natural social interactions, they represent the observations that we will expect to deal with in future works.

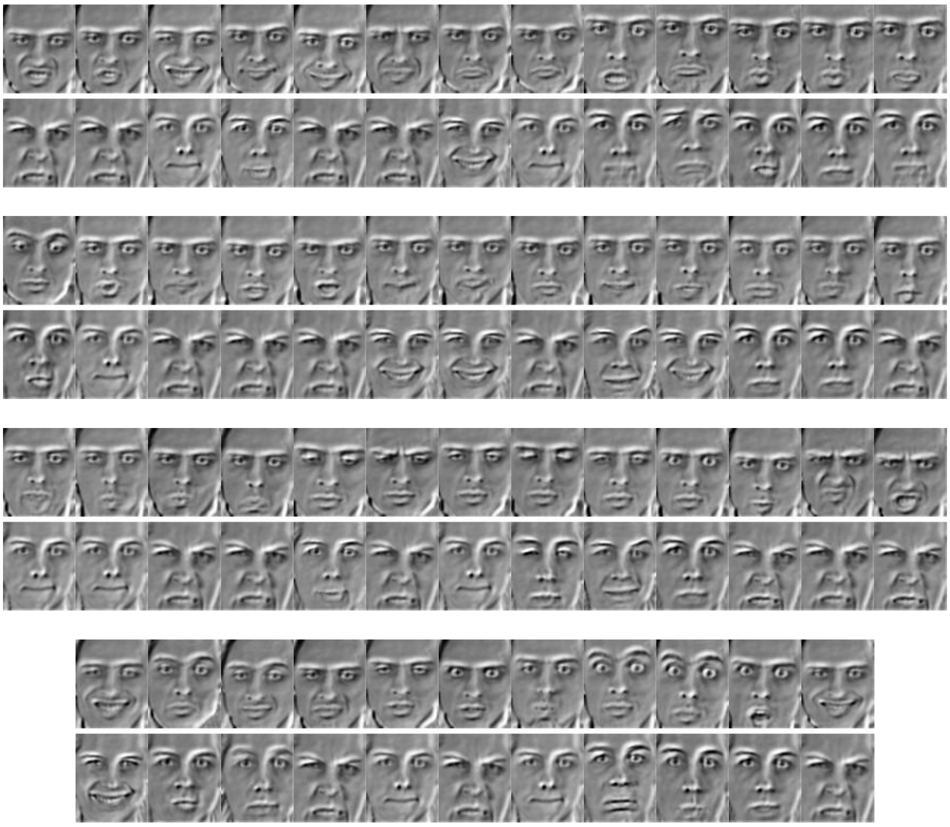


Fig. 5.19: Classification results of Subject1 facial expressions

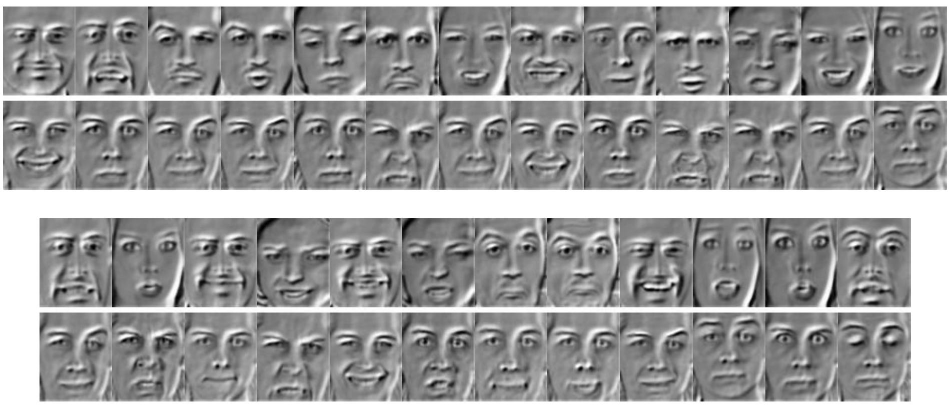


Fig. 5.20: Classification results of faces from the web dataset

Furthermore, we will expect that using faces with a more visible emotional power (unlike the majority of facial expressions in MMI dataset) a latent space with visible axes of core affect variables, like those theorized by Russell, will be generated.

Chapter 6

Conclusions and future works

In this work we have seen how emotions could be used in several useful applications of disparate domains. The majority of these applications need as a prerequisite the not so trivial task of automatic recognition of emotions, as we showed in this work.

Most of the researchers worked on the recognition of the six basic emotions obtaining good results at least in a laboratory asset. Obviously, the recognition of only these few classes leads to poor information for most of applications, especially those concerning a social interaction between a user and a computer.

This issue stimulates us to move into the direction currently taken by most of affective computing researchers, namely the recognition of the continuous core affect space of emotions as supported by Russell theory.

To accomplish this task we presented an architecture for the extraction of faces from videos and a following model for the regression of a latent space. For several reasons, we do not make use of labels and supervised techniques, but we interpreted the regression of the core affect as an unsupervised dimensionality reduction procedure.

Since dimensionality reduction techniques based on linear mapping were not so powerful for our purpose, a probabilistic model based on Gaussian Processes, the GP-LVM, was proposed.

Preliminary tests with this model allow us to investigate the advantages and defects of this tool, obtaining useful guidelines for future improvements.

First of all it can be seen that GP-LVM suffers from differences of identities and poses contained into the dataset used for the training process. Therefore, the dataset used for the training task is crucial to determine the final classification performance of the model. For this reason it is appropriate to use a training set comprehending only a single identity exhibiting a wide range of facial expressions without noise due to different poses.

The process used for the face normalization is quite stable, however not enough to produce a good quality training set. Therefore in future works it will be more appropriate to use frames labelled with at least the position of the eyes and the nose, in order to generate a training set not affected by noise. Nevertheless, the face normalization procedure presented in this work could be useful in future works to capture new observations in the final real life application.

In our opinion, an ideal dataset has to present several subjects involving in a

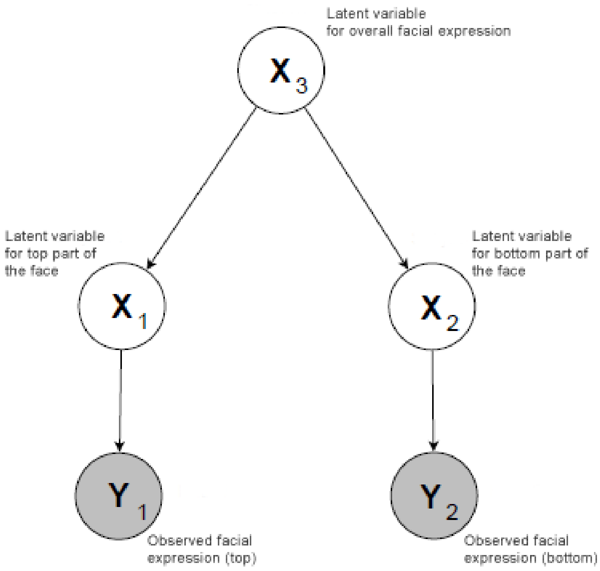


Fig. 6.1: Graphical model of a HGP-LVM for facial expression recognition

variety of believable social interactions in a controlled environment, namely uniform light conditions, absence of occlusions and face with small or null angle of rotations. The best way to collect this kind of data is probably to use good actors/actress and give them a list of plots to act, each of one involving different affective states. Then, as we told previously, each video frame has to be labelled with the position of the facial landmarks.

For our purpose the labelling of this dataset is not necessary, and probably an unnatural process. In fact, currently the only way to objectively describe a facial expression is using AUs; however we saw that producing such a dataset leads to a challenging classification process, due to the small changes among different facial expressions. Unfortunately, the use of an unlabelled dataset implies difficulties during the evaluation process, which could be made only with the use of qualitative evaluations.

In this work we did not contemplate any kind of temporal dynamics on data; however a work of Wang et al. [83] illustrates how to add a temporal constraint on a Gaussian Process, consequently allowing the latent space to include this temporal information. Making use of temporal dynamics means that temporal sequences of facial expressions will generate smoother paths over the latent space, which is a crucial feature for a successive step: the generated paths could be used to classify the affective state of the subject through the characteristics of the path itself.

Conversely to other facial expressions databases, which consider only footages without a history of the affective state of the user, our ideal dataset quoted previously is able to give this crucial temporal information. This cue could be used for example to predict the likely affective state given a set of previous affective states, and can be used for example to simulate a human behaviour in a probabilistic vision.

Other important work that could enhance the results of our model is the Hierarchical GP-LVM by Lawrence and Moore [84]. With this model it is possible to extend

GP-LVM through hierarchies, allowing the expression of conditional independencies in the data as well as in the manifold structure.

In our case it will be possible to consider separately the top and the bottom part of the faces (independency), obtaining a latent space in which more combinations of facial expressions can be generalized with better chances of recognition (Fig. 6.1).

Clearly this work is limited to investigate the affect recognition using only the facial expressions of a subject; however there are also other important modalities for affect recognition, such as gestures, speech, blood pression, ... All these possible modalities can be studied and used togheter in order to enhance the performance of the system.

The overall current classification performance of the model presented in this work is not entirely satisfactory; however it is clear to us the cause of these performance, namely a dataset not suitable for this regression process. In fact it can be seen that GP-LVM is a good tools for dimensionality reduction, but it suffers from noise due to different identities and poses, both of them contained in the datasets used for preliminary tests. Furthermore if the facial muscles movements are not strong enough, the classification of these facial configurations remains challenging, at least using as features the whole pixels of the face. For this reason we are motivated to extend our tests to a different dataset like that presented above, hoping for more satisfactory results.

Aknowledgement

The years during which I studied for the Master include not only good times, but also difficulties that important people have helped to overcome. Thanks to these persons I was able to maintain the right self control in order to fulfil my student duties.

However, conversely to that is usually done in this brief section of the thesis, I do not want to thank these persons, but those that have created difficulties during the achievement of my ambitions.

Thanks to these people I learned that problems can become opportunities, and that it is useless to waste time on choices that simply do not go in the right direction immediately, because sometimes this is a clear signal to support us on choosing other ways.

It is human nature to try to shift the blame on others in order to avoid compromising their social or professional roles. Nevertheless, I learned that a *"Sorry"* can be powerful than the approach previously proposed; in fact it put the other in the position of having to accept your apologies, which is not an easy task for all people. In this way it is possible to understand who deserves our trust and attention and who is not.

From errors it is possible to learn, however it is not possible to change the past. What we can do is to change the present in order to create a more brilliant future. Drastic decisions are sometimes necessary in order to accomplish this task: a change on own lifestyle, a change of friends ... a **change**.

Life is too short to be spent on a single direction. Obviously it is not possible to do all the experiences that our world makes us available, nevertheless it is reductive to focus only on a specific life experience. Probably we are here to learn something; we are not allowed to know what and why, however it is clear to everyone that if we do not make new decisions and we do not try other possibilities our knowledge will remain limited. Although these changes could bring us to errors and challenges, sometimes difficult to overcome, these errors allow us to grow up better.

Anyway, I feel compelled to make a special thanks to Shannon, without which I would not be able to finish in time all those tests presented in this work.

Bibliography

[1] S. Kim and A. McGill. Gaming with mr. slot or gaming the slot machine? power, anthropomorphism, and risk perception. 2011.

[2] N. Epley, A. Waytz, and J.T. Cacioppo. On seeing human: A three-factor theory of anthropomorphism. 114(4):864–886, 2007.

[3] B. Fineman. *Computers as people: human interaction metaphors in human-computer interaction*. Carnegie Mellon University, 2004.

[4] B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Center for the Study of Language and Information, 2003.

[5] R.W. Picard. *Affective Computing*. MIT Press, 1997.

[6] R.E. Cytowic. Synesthesia: a union of the senses. 1989.

[7] J.E. LeDoux. Emotion and the limbic system concept. *Concepts in Neuroscience*, 2:169–199, 1991.

[8] L. Pessoa. On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2):148–158, 2008.

[9] A.R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Gosset/Putnam Press, 1994.

[10] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, January 2009.

[11] T. Fong, I. Nourbakhsh, and K. Dautenhan. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166, 2003.

[12] E. Hudlicka. Affective computing and game design. In *In Proceedings of the 4th Intl. North American Conference on Intelligent Games and Simulation*, pages 5–12, 2008.

[13] J. Sykes and S. Brown. Affective gaming: measuring emotion through the gamepad. In *CHI '03 extended abstracts on Human factors in computing systems*, CHI EA '03, pages 732–733, New York, NY, USA, 2003. ACM.

[14] K. Isbister and P. Doyle. Design and evaluation of embodied conversational agents: A proposed taxonomy. 2002.

[15] B.J. Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do.* Morgan Kaufmann, 2002.

[16] I. Werry, K. Dautenhahn, B. Ogden, and W. Harwin. Can social interaction skills be taught by a social agent? the role of a robotic mediator in autism therapy. In Meurig Beynon, Chrystopher Nehaniv, and Kerstin Dautenhahn, editors, *Cognitive Technology: Instruments of Mind*, volume 2117 of *Lecture Notes in Computer Science*, pages 57–74. Springer Berlin / Heidelberg, 2001.

[17] H. Chein-Chang and C. You Yin. An intelligent fuzzy affective computing system for elderly living alone. In *Hybrid Intelligent Systems, 2009. HIS '09. Ninth International Conference on*, volume 1, pages 293–297, aug. 2009.

[18] S. Thatcher. Reducing aircraft accidents: Can intelligent agent paradigms help? In *Advanced Computer Theory and Engineering, 2008. ICACTE '08. International Conference on*, pages 13–18, dec. 2008.

[19] J.H. Turner. *Human emotions: a sociological theory.* London; New York : Routledge, 2007.

[20] M.D. Klinnert, R.N. Emde, P. Butterfield, and J.J. Campos. Social referencing: The infant's use of emotional signals from a friendly adult with mother present. volume 22, pages 427–432. Developmental Psychology, 1986.

[21] C. Breazeal and B. Scassellati. How to build robots that make friends and influence people. *International Conference on Intelligent Robots and Systems*, 1999.

[22] C. Breazeal, G. Hoffman, and A. Lockerd. Teaching and working with robots as a collaboration. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, 2004.

[23] C. Breazeal and R. Brooks. Robot emotions: A functional perspective. *J.M. Fellous & M. A. Arbib (Eds.)*, 2005.

[24] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life*, 11(1-2):31–62, January 2005.

[25] W. James. What is an Emotion? *Mind*, pages 188–205, 1884.

[26] P.R. Kleinginna and A.M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379, 1981.

[27] J.D. Houwer and D. Hermans. *Cognition & Emotion: Reviews of Current Research and Theories.* Psychology Press, 2010.

[28] A.R. Damasio. Toward a neurobiology of emotion and feeling: Operational concepts and hypotheses. *Neuroscientist*, 1:19–25, Jan 1995.

[29] R.A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1(1):18–37, 2010.

[30] J. Tooby and L. Cosmides. The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology*, 11:375–424, 1990.

[31] S.S. Tomkins. Affect, imagery, consciousness. *The positive affects*, 1, 1962.

[32] C. Darwin. *The expression of the emotions in man and animals*. London, UK: Murray, 1872.

[33] P. Ekman et al. An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200, 1992.

[34] C.E. Izard. Basic emotions, relations amongst emotions and emotion-cognition relations. *Psychological Review*, 99:561–565, 1992.

[35] S. Schachter and J. Singer. Cognitive, social and physiological determinants of emotional state. *Psychological Review*, 69:379–399, 1962.

[36] M.B. Arnold and JA Gasson. *Feelings and emotions as dynamic factors in personality integration*. 1954.

[37] N.H. Frijda and B. Mesquita. *The Analysis of Emotions*. Plenum Publishing Corporation, 1998.

[38] R.S. Lazarus. *Emotion and adaptation*. Oxford New York, 1991.

[39] R.B. Zajonc, P. Pietromonaco, and J. Bargh. Independence and interaction of affect and cognition. *Affect and cognition*, pages 211–227, 1982.

[40] G.H. Bower and J.P. Forgas. *Affect, memory, and social cognition*. Oxford University Press, 2000.

[41] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110:145–172, Jan 2003.

[42] K.A. Lindquist and L.F. Barrett. Constructing emotion. the experience of fear as a conceptual act. *Psychological Science*, 19:898–903, September 2003.

[43] N.R. Carlson. *Physiology of Behavior*. Allyn & Bacon, 2006.

[44] R. Cowie, C. Pelachaud, and P. Petta, editors. *Emotion-Oriented Systems: The Humaine Handbook*. Springer, 2011.

[45] R. E. Kraut and R. E. Johnston. Social and emotional messages of smiling: An ethological approach. (37):1539–1553, 1979.

[46] P. Ekman, E.R. Sorenson, and W.V. Friesen. Pan-cultural elements in facial displays of emotion. 164(3875):86–88, 1969.

[47] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, January 2010.

[48] P. Ekman. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. *Psychological Bulletin*, 115(2):268–287, March 1994.

[49] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834, march 2011.

[50] C.E. Rasmussen and C.K.I. Williams, editors. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[51] C.E. Rasmussen, B.J. de la Cruz, Z. Ghahramani, and D.L. Wild. Modeling and visualizing uncertainty in gene expression clusters using dirichlet process mixtures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):615–628, 10 2009.

[52] M.P. Deisenroth, C.E. Rasmussen, and J. Peters. Gaussian process dynamic programming. *Neurocomputing*, 72(7-9):1508–1524, 3 2009.

[53] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, November 2007.

[54] P. Ekman, W.V. Friesen, and J.C. Hager. Facial action coding system. *A Human Face*, 2002.

[55] M.W. Huang, Z.W. Wang, and Z.L. Ying. A novel method of facial expression recognition based on gplvm plus svm. In *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, pages 916 –919, oct. 2010.

[56] C.H. Ek, P.H.S. Torr, and N.D. Lawrence. Gaussian process latent variable models for human pose estimation. *Machine Learning for Multimodal Interaction*, 4892:131–143, 2008.

[57] N.D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing*, 2003.

[58] I.T. Nabney. Netlab: Algorithms for pattern recognition. *Advances in Pattern Recognition*, 2001.

[59] N.D. Lawrence, M. Seeger, and R. Herbrich. Fast sparse gaussian process methods: The informative vector machine. *Advances in Neural Information Processing Systems*, 15, 2001.

[60] N.D. Lawrence and J.Q. Candela. Local distance preservation in the gp-lvm through back constraints. *International Conference on Machine Learning*, pages 513–520, 2006.

[61] C. Garcia and G. Tziritas. Face detection using quantized skin color regions merging and wavelet packet analysis. *Multimedia, IEEE Transactions on*, 1(3):264–277, sep 1999.

[62] S.K. Singh, D.S. Chauhan, M. Vatsa, and R. Singh. A robust skin color based face detection algorithm. *Tamkang Journal of Science and Engineering*, 6(4):227–234, 2003.

[63] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, March 2007.

[64] H.P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan. Multi-modal system for locating heads and faces. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 88–93, oct 1996.

[65] I. Fasel, B. Fortenberry, and J. Movellan. A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 98(1):182–210, 2005.

[66] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004.

[67] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory*, 1995.

[68] Z. Li Stan and K. Jain Anil, editors. *Handbook of Face Recognition*. Springer, 2011.

[69] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transaction of the ASME – Journal of Basic Engineering*, 1960.

[70] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 1998.

[71] G. Boccignone, P. Campadelli, A. Ferrari, and G. Lipori. Boosted tracking in video. *Signal Processing Letters, IEEE*, 17(2):129–132, 2010.

[72] K.C. Yow and R. Cipolla. Enhancing human face detection using motion and active contours. *Computer Vision ACCV'98*, 1997.

[73] R.L. Hsu and A.K. Jain. Generating discriminating cartoon faces using interacting snakes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[74] T.F. Cootes, C.J. Taylor, and A. Lanitis. Active shape models: Evaluation of a multi-resolution method for improving image search. *5th British Machine Vision Conference*, 1994.

[75] T.F. Cootes and C. Edwards, G. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[76] M. Uřičář, V. Franc, and V. Hlaváč. Detector of facial landmarks learned by the structured output SVM. In Gabriela Csurka and José Braz, editors, *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, volume 1, pages 547–556. SciTePress — Science and Technology Publications, 2012.

[77] I. Tsochantaridis, T. Joachims, T. Hofmann, and H. Altum. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

[78] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, jul 2002.

[79] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *In Proc. AMFG07*, 2007.

[80] R.W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(10):1175–1191, oct 2001.

[81] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, page 5 pp., july 2005.

[82] R. Urtasun, D.J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 403–410, oct. 2005.

[83] J.M. Wang, D.J. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *In NIPS*, pages 1441–1448. MIT Press, 2006.

[84] N.D. Lawrence and A.J. Moore. Hierarchical gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 481–488. ACM, 2007.